ASSESSING INTERACTIONAL COMPETENCE IN A MULTIPARTY ROLEPLAY TASK: A MIXED-METHODS STUDY

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAIʻI AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

SECOND LANGUAGE STUDIES

NOVEMBER 2018

By

Patharaorn Patharakorn

Dissertation Committee:

James Dean Brown, Chairperson
Gabriele Kasper
Thom Hudson
Betsy Gilliland
Seongah Im

ACKNOWLEDGEMENTS

I would like to thank all my committee members, Dr. Brown, Dr. Kasper, Dr. Gilliland, Dr. Im, and Dr. Hudson, without whom this dissertation would not have been remotely possible. This statement is such a cliché. But as I reflect on my six years here at UH, I cannot help but feel tremendously fortunate, and so grateful, that I have had such strong support from each of the committee members. More than my profound respect for their academic expertise in their respective fields is my admiration for their passion and dedication for teaching and mentorship. I am thankful to have learned from all of them about the kind of person I would like to one day be for my students.

I would also like to thank Chulalongkorn University Language Institute, who had provided financial security for me during the first four years of my education here, and also for allowing me the time away from my teaching responsibility to pursue this personal and academic growth. This dissertation was also possible thanks to the generous funds from Small Grants for Doctoral Research in Second or Foreign Language Assessment from ETS. Their support was essential during the data collections of this study.

Finally, I would like to thank my parents, my family, my husband and my friends for all their unconditional love, support, reassurance, and encouragement.

ABSTRACT

In an effort to develop an assessment instrument in measuring interactional competence (IC) with a method that is congruent with the current research findings on IC and IC development (e.g., Hall, Hellermann, & Pekarek Doehler, 2011; Pekarek Doehler & Pochon-Berger, 2015), the present study investigated students' performances on a multiparty roleplay on a task called Socializing. Using the sequential mixed methods design (Greene, 2007; Tashakkori & Teddlie, 2003), the study explored empirical evidence garnered through qualitative and quantitative research methods to test if the proposed rubric can provide a valid and reliable measurement of IC on this performance assessment task.

The participants of this study were 180 undergraduate engineering students at a university in Thailand who were taking an EFL course that targets social communication skills in professional contexts. Students were randomly grouped together and were asked to have a conversation for 10 minutes, in a roleplay task which they must introduce themselves as their character and try establishing business contacts for their hypothetical companies. The data for this study included 34 video-recordings of the group roleplay performances.

Conversation analysis (Clift, 2016; Sacks, 1992; Schegloff, 2007; Sidnell & Stivers, 2013) was employed to identify comparable interactional activities and determine the interactional methods students utilized in carrying out those activities. The productive activities are self-introduction, work talk, business contact exchange, post-conference arrangement talk, and an interaction to bring about the termination of the roleplay performance. Three recipient actions include students' management and display of their understanding, students' management of alignment, and finally, their display of affiliative stance.

Six raters from various teaching and training backgrounds were recruited to apply the proposed rubric in evaluating the students' IC on the eight items, combining both productive and

recipient actions. The Many-Faceted Rasch Measurement (Linacre, 1989) with the Partial Credit Scoring model (Masters, 1982) provided integrated measurement reports of the rating practice. The findings revealed that students' ability on this IC construct mostly exceed the difficulty of the *socializing task*. Self-introduction and understanding display had been identified as the two easiest items, followed by alignment display, work talk, affiliation display, activity termination, making post-conference arrangements, and bringing up contact exchange, respectively. The analysis also suggested that most raters were reliable in applying the rating scale, though they demonstrated a higher degree of uniformity in evaluating productive activities compared to their ratings of recipient actions. Overall, the mixed methods research design is seen to have provided a much-needed framework in this process of exploring the validity evidence of the proposed rubric in assessing IC for the multiparty roleplay performances on the *socializing task*.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Talk is a site where individuals come together to articulate and manage their collective histories via their resources (Hall, 1995). The ability to participate competently in talks and conversations is one of the most fundamental goals of learning an additional language. Derived from communicative competence (Bachman, 1990; Canale, 1983; Canale & Swain, 1980), interactional competence (IC) is commonly understood as the ability to interact naturally and appropriately in social situations. The definitions being used and referred to as IC seem to have evolved over the years. When Kramsch (1986) proposed the concept of IC as part of her argument for a redirection of language teaching goals from 'proficiency' to 'interactional competence,' she urged that language teaching practices should move away from focusing solely on linguistic accuracy in favor of fostering IC - the teaching and learning of 'interactional processes' and discursive skills in order to cultivate learners' intercultural awareness in cross-cultural interactions (p. 370). Since then, the study of IC had flourished in the field of language testing and assessment, led by the work of Young and He (1998) and McNamara (1997). The interests in IC among language testers coincides with the emerging research interests among conversation analysts in second language learners' interactional practices (i.e., Firth & Wagner, 1997; Hall, 1995)

However, when it comes to how second language spoken discourse is evaluated, this ability to interact competently is still undertheorized in language assessment. Most traditional rating scales tend to focus on linguistic skills and their fluency of delivering the message, which comprise two major aspects of test takers' ability, accuracy and fluency, when oral performances are being assessed (Lazaraton, 2014). These traditional linguistic skills refer to criteria such as

grammatical accuracy and complexity, pronunciation, and vocabulary. The fluency criterion usually refers to the 'flow' of delivery, and could include consideration of cohesion and coherence of speech. These criteria do not sufficiently capture skills required in interactive communication because they lack a sensitivity to "inter-" individual skills that an examinee displays in relation to his or her co-participants. In other words, with this traditional approach, the occasion of talk and interaction in speaking test formats is merely used as a stage for assessing other constructs. The skills which go into enabling talk-in-interaction to take place have been mainly overlooked or underrepresented by performance rubrics and traditional test construct definitions.

While many of the go-to criteria in assessing a spoken discourse have been directly borrowed from assessing written discourse (i.e., grammatical accuracy, vocabulary, or coherence), other criteria which are unique to assessing oral production mostly reflect the time-sensitive nature of the spoken discourse. Language testing may have evolved far enough to be able to score the “fluency” construct consistently by measuring pauses, speech rates, or a count of dysfluency markers (Fulcher, 2015). However, as we can see that there is a vast difference between "fluency" in a monologue and "fluency" in a normal conversation, this fixed view of "fluency", and the fact that language testers have gotten more consistent in recognizing it as such may actually be problematic.

In a recent movement in language testing, attention has turned to the construct of interactional competence (IC) (e.g., Al-Gahtani & Roever, 2012; Roever & Kasper, 2018; Youn, 2015), providing a possible framework which can be used to look at interactional skills in spoken discourse in its own right. This line of research inquiry shares a goal to help validate the assessment practices that wish to say something about test takers’ ability to conduct themselves in normal conversations in their second languages.

Background: A Look Back at IC in Language Testing and Assessment

In the earlier form of incorporating social constructs into the assessment of communicative competence (Bachman, 1990; Canale & Swain, 1980), much work had been done on assessing pragmatic competence (Brown, J. D., 2001; Brown, J. D. & Ahn, 2011; Grabowski, 2013; Hudson, 2001; Hudson, Detmer, & Brown, 1992, 1995; Kasper & Ross, 2013; Roever, 2011; Yamashita, 2008; Youn, 2013) before it was extended to the more encompassing construct of IC.

Attempts in creating an assessment of IC, however, have been scarce, and we see more use of interactional ability criteria as part of the rubrics which are used for assessing communicative language competence in classroom contexts and some specific occupational contexts. Quite likely, the notion that IC is fundamentally situated and co-constructed (Young, 2011; Young & He, 1998) had made matters surrounding assessing IC complicated.

Many studies, however, have shed light on the importance of possible interactions between raters, scoring rubrics, and certain interactional features of student performances. Douglas (1994) found little relationship between scores on a test and the language actually produced by test takers. He discussed that it is possible that raters might have been influenced by aspects of language performance that have to do with communicative language ability that was not present in the scoring criteria. More association between scores and performance was found in another study into the IELTS speaking test. In comparing high and low scoring test takers, Seedhouse (2012) found that high scoring students systematically have interactional ability to develop topics more extensively in interaction, display engagement in their participation, and construct their a 'professional' and 'internationally oriented' identity through their talk. May (2011) has taken the rater's perspective to see what interactional features are salient to them, pointing to future research directions compatible with that of Seedhouse (2012) and Douglas

(1994) that rubrics with explicit data-derived criteria for IC are needed for assessing oral communicative performance. At the very least, these findings point to a strong susceptibility that raters implicitly have towards IC, and it is with this belief that this current study is attempting to make IC more explicit.

Overview

The important question of "whose competence?" asked by McNamara (1997) in reference to the co-contributing nature of social actions, remains open for further exploration. Given the institutional use of most language tests is to provide information about an individual language learner's knowledge and ability, it is useful to have a model which can help untangle individual contributions from what is collectively accomplished. One of the main tasks at hand is to establish what observable attributes for interactional competence look like in second language assessment, so that we can optimize the tasks and rating instruments to better suit the aspects of IC that second language users and test takers can display.

In order to explore the construct of IC in oral assessment tasks, IC addressed in this study is situated in an elicited talk in a group roleplay task. The task chosen as the target of this study was designed to assess socializing skills in a work-related informal situation when English was used as a lingua franca. This group roleplay format was chosen because it involves unscripted talk among four to six participants who have individual and collective goals they have to achieve in the roleplay. Also, it is this talk-in-interaction element that would allow for plenty of opportunities to inspect how IC can play a role in determining the success or failure of performance in this task.

This study's aims are first to make explicit the construct of IC displayed in this specific context and task. Then, a performance data-driven rubric for assessing those aspects of IC will be constructed and checked for its reliability and validity.

To address this co-construction notion within IC, this study adopts the approach of IC development studies (Nguyen, 2012a; Pekarek Doehler & Pochon-Berger, 2015; Pekarek Doehler, Wagner, & González-Martinez, 2018) that argues we can examine an individual's competence display while paying attention to how the competence is co-constructed by participants in interaction. In other words, this study takes a slightly different viewpoint from the mainstream view of IC in language assessment (more discussion on this view in Chapter 2), as it seeks to provide evidence to support that, in a given specific set of actions and situations, individuals can show themselves to be more or less interactionally competent, and this could be viewed as the person's interactional competence within that co-construction (Kasper & Ross, 2013).

This study reports the development process of a data-driven rating scale which targets the construct of IC situated in a multiparty open roleplay task taken from an existing classroom assessment task used since 2012 called a *socializing task*.

The literature review is divided into two chapters. Chapter 2 provides a summary of theoretical frameworks including the literature on IC and IC development, IC as a construct in language assessment, and the validity and validation framework for assessing language performance. Chapter 3 addresses as the research framework the mixed methodsapproach adopted in this study. This chapter discusses mixed methodsresearch designs, its strengths and potential weaknesses, and the challenges in combining certain qualitative and quantitative research methods. Additionally, Chapter 3 also reviews the literature surrounding performance

assessment, including the issues of raters' reliability, rubric construction and application, and rating scale development and validation procedures.

After reviewing the theoretical and research methodological frameworks, Chapter 4 presents the research questions this study is designed to investigate. The chapter also provides a detail description of the current study's sequential mixed methods design, the data collection methods, and data analysis procedures. Given the sequential mixed methods design of the study, the study divides the result and discussion chapters into two parts. Chapter 5 reports the qualitative findings from student performance data, and Chapter 6 provides a discussion of the findings in relation to the corresponding research questions. Then, the quantitative results from implementing the proposed rubric for assessing IC are reported and discussed in Chapters 7 and 8. Lastly, the final chapter concludes the study by discussing its limitations, potential implications, and suggestions for future research.

CHAPTER 2

THEORETICAL FRAMEWORK

To provide the theoretical underpinnings for this study, three areas of literature are covered in this chapter. First, the bulk of this chapter will review the currently available work on the construct of interactional competence and interactional competence development in L2 learners. Second, the chapter will then discuss the treatment of IC in the field of language assessment, and finally, the chapter will provide a summary of test validity and validation framework concerning the assessment of L2 performance.

Interactional Competence and Interactional Competence Development

To develop an assessment of interactional competence (IC) in second language learners, language testing researchers can draw from the cumulative body of research on L2 interactional competence and interactional competence development. This area of research interests has been studied by conversation analysis (CA) researchers, whose methodological orientation is informed by the long tradition of sociology's ethnomethodology (Garfinkel, 1967). Because CA pays close attention to conversation participants' micro-management of local semiotic resources, it can provide a rigorous analytical toolkit for studying IC as it enables researchers to capture moment-by-moment interactional work with an intricate level of detail. In this section, the structural organizations of interaction, as described in foundational CA studies, will be discussed before the chapter provides a summary of findings pertaining to second language learners' IC and IC development based on ethnomethodological conversation analysis (EMCA) research. Then, the status of interactional competence as an assessment construct in the field of language assessment will be reviewed and discussed.

The Order of Interaction

Conversation Analysis (CA) research has taken an interest in describing in detail the procedural infrastructure of interaction – the way in which competent members of a speech community manage their talk. Following Garfinkel's ethnomethodology in describing social interaction, it is important to note that researchers should take the emic, or the participants', perspective in analyzing action-in-interaction as it was produced and oriented to in the first place by the participants for the co-participants, not the analysts (Schegloff & Sacks, 1973). CA investigates the day-to-day conversations under the idea that there are orders at all points in interaction, and it is this structural procedure which members socially share that enables them to communicate with one another (Sacks, 1992; Sacks, Schegloff, & Jefferson, 1974). CA treats occurrences of conversation or talk-in-interaction as achievements that both the speaker and listener co-construct together in a moment-by-moment sequential organization. In order to be able to describe IC later in the chapter, some brief reviews of CA's concepts which constitute the building blocks of social interaction are necessary.

Turn-taking system. Central to the participation methods of talk-in-interaction, Sacks et al. (1974) proposed that conversations are composed of turns, and that turns appear in an emergent sequential structure. As a building block in conversation, a turn is a package where an action is implemented, and this is also referred to as turn-constructional unit or TCU (Schegloff, 2007) which can be realized in a form or sentences, clauses, phrases, or lexical items. Rather than adhering to any linguistic properties of an utterance, talk-in-interaction is organized and regulated by the participants one TCU at a time (see Clift, 2016 for discussion on how grammar is configured in interaction). In taking turns at talk, we are constantly oriented to a possible completion of an ongoing TCU as a transition relevant point (TRP) to issue the next TCU. A critical feature of a turn is that it constitutes an action recognizable in and to that particular

context and co-participants. Linguistic and grammatical resources are part of the ingredients in producing and interpreting TCUs, but so are other phonetic and non-verbal resources (Clift, 2016). In general, speakers take turns in producing one TCU at a time. In cases where the current speaker needs more than one turn to talk, for example in the case of story-telling, then some interactional work is required to project that a multi-turn unit is forth-coming (e.g., "*Did you hear about …?"*), so that the recipient would hold off taking their turn to make room for such units. From this scenario, projecting a multi-turn unit and recognizing appropriate points at which to withhold a turn are interactional accomplishments initiated by one participant and co-constructed into being with the other co-participants.

When the goals of the interaction are observable either from within the interaction or imposed upon the participants by their institutional goals (Sanders, 2003), inferrentially observable still in the interaction, we should then be able to judge the degree of accomplishment given the design and composition within their turn construction units (TCUs).

Sequence organization. CA is particularly concerned with how utterances can accomplish a certain action not only because of their designs but also their placements. The current action projects what is the relevant action in the next adjacent turn. In the same vein, an utterance is, therefore, interpretable in relation to what immediately precedes it. In analyzing conversational data, CA focuses on 'adjacency pairs' as a unit of sequence construction (Schegloff, 2007; Schegloff & Sacks, 1973), or sequence co-construction to be more precise. Many conversational actions occur in pairs, such as question-answer, request-grant/refusal, or even a *hello-hello* greeting exchange. It is important to note, however, that the first action makes relevant the next action, but it does not determine what must come next (Kasper & Ross, 2013). For example, when someone understands a turn being directed at themselves, they may display their understanding by nodding, uttering some tokens like '*uh-huh*' (Schegloff, 1982a) or

'*oh!*'(Heritage, 1984a) or '*I see*,' or they may formulate a response that showcases their understanding as they come up with relevant responding actions. It should be noted that these different tokens do not just signify understanding; they can carry out different social actions that can be later observed in the on-going streams of interaction.

Adjacency pair structure is found to be very stable and can be applied in describing many types of sequence organizations, from very straightforward sequences to much more complex, subtle, and nuanced conversational activities (Heritage, 1984b), e.g., confirming allusions (Schegloff, 1996). With adjacency pair construction, actions in the first and second pair parts are mutually accountable. When the relevant or expected 'next' action does occur, it is characteristically treated as normal and requires no special explanation. On the contrary, when an expected or relevant 'next' action does not occur or is not forthcoming, the breaching party is then held accountable, with some form of explanations or accounts are then expected (see Schegloff, 2007 for in-depth explanations).

The point here is that these are some of the normative ways of making understanding visible to others; because it is normative, it is recognizable by the interlocutors, and possibly also analysts if they share the membership knowledge of the speakers. With that said, however, members' conduct can deviate from the norms. When one party in a conversation behaves differently from the norms, there are observable consequences that both parties may have to deal with through initiating repair sequences until things normalize again. What is treated as normal by the participants is observable to researchers because even when the norms are violated, it is still noticeable as participants would then provide an account for those anomalies in their interactions.

Sequential organization or overall structural organization. Similar to the sequence organization of turns, when one sequence is brought to a close, another sequence can then be

launched next with or without some transition activities in between. Schegloff (2007) refers to the overall sequential organization as the sequences of sequences. In ordinary conversations, the next sequence can be formulated as a reciprocal sequence in the same activity-type series, conducting the same action in reverse direction, i.e., "*how about you?*" (Schegloff, 1986). Another kind of relationship between sequences can be more tight-knit so that they make up a larger course of action in successive parts. Each sequence implements the next stage in the course of action (Schegloff, 2007). Jefferson considered the sequential organization of this kind a "socially organized ‘package' which contains standard components in a standard order of occurrence" (Jefferson, 1988, p. 418 emphasis in the original).

Describing multiple sequences of action grouped together with an overall structural organization is an open-ended endeavor for conversation analyst researchers in identifying recurring practices regularly found across a given corpus (i.e., Jefferson, 1988). For example, an episode of talk during an opening of a birthday present would involve sequences from acknowledging the giver, reading the card, opening the gift, positively assessing the gift, and then finally thanking the giver (Good & Beach, 2005). In another example from a more professional discourse, an episode of consultation during a medical visit at a doctor’s office would involve sequences from presenting the concern, gathering information, diagnosing, until treating the concern (Robinson, 2003; Robinson & Heritage, 2006).

The shape and order of the course of action form a supra-sequential coherence which Sacks (1992) called a “big package” (p. 354) or a set of pre-organized sequences (p. 355). The coherence derived from the overall structural organization can help explain the conceptualization of *activity* as a unit of interaction (Robinson, 2013) or an interactive practice (Hall, 1995). An important methodological point in analyzing interactional activities is that CA takes an emic analytical perspective; thus, describing overall structural organization must focus on how

participants orient to such organization as normative, coherent, and as something that once they 'departed from,' they must 'return to' (Heritage & Sorjonen, 1994). Robinson (2013) noted from the longstanding research on institutional talk that the studies of overall structural organization are more commonly reported in institutional interactions, (e.g., Drew & Heritage, 1992; Heritage & Clayman, 2010) given their often explicit institutional goals compared to those of ordinary conversations. This line of study of institutional interactions tends to invoke the interactional rights, expectations, and obligations associated with such identities from within the interaction, and thus can render a great help in informing the standard setting in language assessment practices.

IC and IC Development

A number of studies in recent decades have discussed the nature of IC (e.g., Hall, 1995; Hall et al., 2011; Kasper, 2009; Kasper & Ross, 2013; Nguyen, 2012a). Hall (1995) used the term IC to refer to the ability to participate competently in interactive practices. In her study of teacher and student interactional competencies in classroom interaction, competent management included an ability to develop and manage topical talk, which entails both an ability to construct utterances and an ability to figure out what is going on topically. In the first edited volume on the topic of IC, Hall et al. (2011) proposed that IC can be viewed in two parts: the knowledge of relevant available resources and the ability to make use of them to carry out interactional work in ongoing talks. The first part subsumes under the second as IC is only observable in interaction (Mehan, 1979).

The tenets of IC are operationalized to mirror the apparatus of CA. This entails an ability to competently and appropriately conduct turn-taking, sequence organization, repair practices, and topic management, and also invoke relevant participation frameworks, etc. These kinds of

abilities tend to be implicit among competent members of any speech community. However, for interactions which involve parties who are not-yet-competent speakers such as second language learners, the cumulative body of work that describes their L2 interactional competencies can be found in ethnomethodology-oriented second language acquisition studies using conversation analysis (CA) as their research method (CA-SLA for short) (i.e., Brouwer & Wagner, 2004; Gardner & Wagner, 2004).

Kasper and Wagner (2011) point out that CA-SLA work is doubly concerned with IC. On the one hand, it seeks to explain the kinds of competence that enable learners to acquire their L2 either inside or outside the classrooms, and on the other hand, it studies the L2 speakers' IC when IC is the learning target in itself. The former area of research has been instrumental in describing in detail the occasions in which L2 learners display their IC through mobilizing their available resources and accomplishing social actions in interaction with their co-participants (i.e., Lee, Y.-A., 2006; Theodórsdóttir, 2011). In the latter area of research, studying developmental stages of L2 IC poses a set of challenges for CA-SLA researchers given that CA does not normally make analytical claims beyond what is observable in the here and now, so researchers are faced with an inherent challenge of how to deal with the concept of development and identify the process in which development of IC can be observable.

The challenges are neatly summarized by Pekarek Doehler and Wagner (2010), who pointed out that development in the form of observable differences between time A and time B can be due to changes in the local context instead of changes due to development, or in the cases of cross-sectional research, due to their differed levels of IC. In conducting comparative research in CA, either longitudinal studies or cross-sectional studies, analysts are confronted with two methodological challenges: first, how to warrant comparability across their collection, and second, how the changes or differences in interactional practices can be documented and

showcased (Pekarek Doehler & Berger, 2018). In other words, the challenge facing researchers is that they need to establish commonalities among interactional occurrences, enough that they can be categorized as doing the same action. They also need to establish differences among the occurrences enough that their claim of change or development is then warranted.

Despite these challenges, several significant contributions capturing different dimensions of IC development among L2 learners have been made. These studies typically traced a developmental trajectory of a distinct action or course of action; for example, turn-taking (Cekaite, 2007) or storytelling (Pekarek Doehler & Berger, 2016). Longitudinal studies of IC development contribute to our understanding of how learners' methods of accomplishing an action or a course of action change or develop over time, while cross-sectional studies offer evidence as to how learners at different proficiency levels differ in the ways they accomplish an action or a course of action. Discussing the existing work on L2 IC development, the findings below are grouped based on their different analytical targets, starting from interactional mechanisms of turn-taking practice and repair practice, to sequence organizations of actions and course of actions.

Turn-taking practice. In a study which traced IC development of a Kurdish child's self-selection practice in doing turn-taking during classroom interactions in a Swedish L2 context, Cekaite (2007) demonstrated how the self-selecting practice had progressed from non-participation to disruptive self-selection through non-traditional techniques to eventually conducting self-selection in conventionally appropriate ways. The study showed that later in the data collection, the child could issue self-selections at non-disruptive and sequentially appropriate positions showing an awareness of TRPs or TCU boundaries. Moreover, the designs of the child's self-selection were also found to have improved in the topical relevance with ongoing talk. This also shows how the teacher had over time oriented to the child's self-selection

practices as more appropriate, which in turn shaped the child's interactional practice to develop in this direction.

Repair practices. CA's *repair* refers to the class of treatments needed when interactional troubles arise. Troubles can be anything which participants judge as impeding their communication. The one who initiates the repair is the one who perceives it as repairable. This is a very important competence as it is a vital mechanism for the participants to address and resolve any troubles in speaking, hearing, and understanding (Schegloff, Jefferson, & Sacks, 1977).

In Hellermann's (2011) large-scale project he studied how language learners accomplished repairs when they just communicated among themselves, in comparison to how repairs are accomplished in talk between language learners and native speakers. He found that language learners do produce repair regularly in the same location to the native speakers' norm. He was able to show that the 'next turn repair-initiation' could be fundamental to human interactional culture (p. 166), regardless of language. He also found that learners at a later point use a wider repertoire for doing other-initiated repair, like using embodied actions, open class initiators (Drew, 1997), and more specific repair initiators; the learners were able to treat a variety of objects as repairable in talk, such as discourse structure or a course of interactional actions.

Sequence organizations. Sequence organization refers to a series of turns that are normatively organized together to enact courses of meaningful actions in interaction (Schegloff, 2007). Special attention has been given to identifying L2 learners' IC development in initiating an action or opening a course of action. Pekarek Doehler and Pochon-Berger (2015) highlighted some crucial requirements on the part of the speakers as they are initiating a sequence or managing an opening sequence of a course of action. First, the speaker needs to display to their co-participants what actions or what course of actions are being opened so that the actions are

recognizable. Also, the speaker needs to display the local relevancy of the upcoming course of actions to the ongoing interaction so that the actions are acceptable for the co-participants. To meet these requirements, speakers must *recipient-design* or customize their turns to fit situations and the co-participants present at the time of talk.

The action which has been studied the most is the opening of a storytelling sequence (Jefferson, 1978; Mandelbaum, 2013; Sacks, 1972). As we briefly mentioned earlier, storytelling sequence is rather special in that in order to launch the sequence, the normal turn-taking mechanism must be put on hold to make room for one party to take an extended amount of turns while the others take a back seat as they participate in storytelling talk (Sacks, 1992). In a study which compared unsolicited story opening practices among ESL students at the beginner and intermediate levels, Hellermann (2008) reported that beginner learners are less successful than intermediate learners at displaying the local relevancy of their story to the ongoing talk in launching the story. Beginner level students also tend to skip the pre-storytelling sequence which has a crucial function of signaling their co-participants of the forthcoming change in their turn-taking system and helps facilitate such transition. The difference between the two levels in Hellermann's study is an ability to recipient-design their storytelling opening. This also corroborates with another study of a larger corpus by Lee and Hellermann (2014) which also reported similar results.

Pekarek Doehler and Berger (2016) conducted a longitudinal study tracking the development of Julie, an advanced L2 French speaker from Germany, while she stayed with a host family in France over the period of nine months. The researchers demonstrated that while Julie was able to do some prefatory work in launching her storytelling at the beginning of the study, towards the end of her stay, her story prefacing sequence became more extensive and better recipient-designed. This included her control over a larger variety of techniques to display

local relevancy of the upcoming story with the ongoing talk and to signal to her co-participants what kinds of story were about to be underway to aid their reception and expectation of the story.

Another type of action which has been researched for IC development in sequence initiation is students negotiating task opening. Hellermann (2007, 2008) reported a similar trajectory in that beginning level students' task openings are most of the time abrupt, showing no prefatory work with minimal to no recipient design to ensure recipiency and mutual recognition of the task opening before progressed to task accomplishment. On the other hand, intermediate learners were more able to engage in task opening talk and showed greater ability to recipient-design their talk.

In another type of action, IC development in transitioning from one topic in conversation to another has also been studied. Before we proceed, it should be noted that CA treats the notion of 'topic' differently compared to what is commonly understood of topic organization. CA does not focus on what the topic is about, but rather what the participants treat as a topic during talk-in-interaction. Schegloff (2007) suggests that it is better to examine topic with respect to action than with respect to topicality, as treating topic in its topicality has shown to be more complicated than it appeared to be (see Schegloff, 1990 for more detailed discussion). Topic management is, therefore, a kind of interactional accomplishment of a series of sequence organizations: topic initiation, topic maintenance, topic shift, and topic closure.

Lee and Hellermann (2014) studied topic shifting practice among second language learners of English in South Korea. This longitudinal data is part of a 10-month weekly speech practice. In each session, members take turns to produce a five-to-fifteen-minute long presentation. The study featured one of the students, tracing how this speaker handled topic shift during her storytelling. In this study, Lee and Hellermann (2014) showed that the focal speaker was able to improve their change of topic from an abrupt topic shift at the beginning of the study

to being able to signal a topic shift appropriately by employing topic shift markers to coordinate with the recipient that of what is to come.

In a cross-sectional study which compared the performances of learners at different proficiency levels in initiating requests in roleplays, Al-Gahtani and Roever (2012, 2013) demonstrated that while the higher proficiency group managed their requests using the methods similar to the L1 findings, the lower proficiency learners were more likely to launch their request with minimal to no pre-request sequence. To establish what typical competent interactions look like, Excerpt 2.1 below is reproduced from Schegloff (2007).

Excerpt 2.1 *Request Sequence Reproduced from Schegloff (2007, p. 47)*

```
1      Bon: Fpre →   But- (1.0) Wouldju do me a favor? Heheh
2      Jim: Spre →   e(hh) depends on the favor::, go ahead,
3      Bon:          Didjer mom tell you I called the other day?
4      Jim:          No she didn't.
5                    (0.5)
6      Bon:          Well I called. (.) [hhh ]
7      Jim:                             [Uhuh]
8                    (0.5)
9      Bon: Fb →     .hhh 'n I was wondering if you'd let me borrow
10                   your gun.
```

The organization of a request sequence is one of many activities very well documented in CA literature. In English, there is some preliminary interactional work that speakers would normally attend to before one party launches a request. This type of turns is functioned as *pre-sequence* as the action carried out in such sequence is to obtain the legitimacy for the main action to be launched later. In the above excerpt, Bonnie issued a pre-request in line 1, "*would you do me a favor?*", to check if she could have Jim's permission to launch a request. In line 2, we can see that Jim gave a go-ahead signal while being explicit that his permission is only for Bonnie to launch her request, and not that he agreed to the unspecified favor she was asking. Bonnie's actual request came much later in lines 9-10, finally revealing that she wanted to borrow Jim's gun. By deferring her request with multiple hearable breathings and other delays after the pre-request has been ratified, Bonnie displayed that her request was carrying a high degree of imposition necessary of more mitigation.

In contrast to the example above from L1 data, Excerpt 5.2 reproduced from Al-Gahtani and Roever (2011, p. 58) below illustrates a request produced by a low proficiency learner. The sequence organization of this request initiation is problematic because the talk progressed from greeting sequence in lines 1-2 to P's request in line 3 lacking any preliminary exchange normatively found in competent request formulation.

Excerpt 2.2 *Lecture Notes, Beginner*

```
1      P:    hi teacher
2      I:    hello:: ((name))
3 →    P:    this me (.) I want paper (.) my cla::sses (.) OK?
```

In a series of studies by Al-Gahtani and Roever (2012, 2013, 2018), which analyzed how students from different proficiency levels carried out requests and refusals, they found that an ability to employ prefatory work in conducting their requests and refusals increases as they move up the proficiency levels from little to no use of pre-sequence as the above example illustrated to a much more recipient-designed turn formats employing lexical and sequential resources much closer to what has been reported on L1 data.

The summary so far has focused on the findings of IC and IC development particularly of L2 learners. However, some studies have also documented IC development of L1 speakers during their professional training. Rine and Hall (2011) discussed how international teaching assistants demonstrated their ability to more appropriately invoke participation frameworks to be recognized as competent teachers through orienting to “teacher space” and become more teacher-like in the ways they started and ended their lessons. Nguyen (2011) looked at pharmacist-patient consultations and studied the pharmaceutical interns' IC development. In an advice-giving sequence during the consultation, Nguyen discussed the interns' increased ability to design their explanations in a way that is easier for their recipients to understand, making their consultation more effective. Later in the program, the intern was able to produce advice that uses less

technical terms and more specific to how the patients can observe his or her symptoms for when there is an allergy.

This dimension of IC relates to how relevant roles, stances, and identities are being appropriately invoked in talk. This layer of work is highly pervasive, and it could equally be communicatively obligated, as well as communicatively strategized. Role relationships that we often encounter are, for example, teacher-student, parent-child, expert-layperson. Realizing these roles, as well as the socially accepted stances and identities, takes interactional work. Hence, one has to be interactionally competent in being able to do so appropriately.

Taking all things together, we can see that IC encompasses more than the control over linguistic or grammatical resources. It involves the abilities to maintain and coordinate social interactions with co-participants, to invoke relevant identities through talk-in-interaction, and to repair the course of action when any troubles in interaction arise. With the growing body of findings on IC development, the trajectory generally identified as more interactionally competent is quite uniform in pointing towards accomplishments of actions that are more efficient and resourceful.

This take on second language acquisition focusing on IC development is worth pursuing for a number of reasons. First, it puts the notion of language use at the center of the language learning agenda. Brouwer and Wagner (2004) highlighted this alternative view as they noted that an account of language learning cannot just pay attention to the formal linguistic items as in psycholinguistically oriented SLA work (also see Atkinson, 2011). We must recognize interactional skills and interactional resources at all points of the learning experience and study how the L2 speakers construct their actions and make sense of their world as they participate in their L2 discourse community. Moreover, the learners are no longer viewed as handicapped language users (Veronis & Gass, 1985), or inferior to their native speaker counterpart in any

permanent ways. With this view on language learning, learning to participate in an interactionally competent manner is not limited to only second language speakers, but includes any speakers who are entering a new discursive practice of which they are not a member. A discourse practice can be defined at a professional level, community level, or any specialized form of talk.

IC in Language Assessment

The construct of IC has been given a rather unique status in language testing and assessment compared to other assessment constructs. Older key papers on IC in the field of language assessment argued that IC is fundamentally co-constructed, local, and situated (McNamara, 1997; Young, 2000; Young & He, 1998). Some skepticism had been expressed about the possibility and practicality of measuring the construct of IC in language assessment (Fulcher, 2010), but the field of language testing and assessment was reluctant and cautious as to how the construct of IC should be treated and operationalized. Although having voiced the importance for the language testing field to operationalize and interpret IC for individual learners, McNamara (1997) pointed out that this need will not be easily fulfilled given the inherently social nature of IC.

A small movement within the field of language assessment attempted to cope with this social construct – to untangle this notion of co-construction-- by treating it as variables (e.g., Bonk & Van Moere, 2004; Brown, A., 2003, 2005; Davis, 2009; Galaczi, 2014; Gan, 2010; O'sullivan, 2002; Ockey, 2009). Nakatsuhara (2013) studied factors of learner characteristics to see how individual test takers affect the co-construction of group oral tests, and May (2011) took the raters' perspective to see what interactional features are salient to them, pointing to future

research directions that rubrics with explicit data-derived criteria for IC are needed for assessing oral communicative performance.

While attempts in teasing out co-participants' contribution to the overall shape and patterns of conversation can help language testers better control for irrelevant variables from having exerted unwanted biases in the process of performing or rating interactional data, it has not provided many answers as to what the ability to interact in social situations actually look like and where can such IC constructs are located in individual learners.

There is a strong trend of language testers bandwagoning on identifying "features" of interaction which are perceived as displaying mutual accomplishments. Galaczi (2008) conceptualized patterns of co-construction into four styles: collaborative, parallel, asymmetry and blend, based on dimensions of mutuality, equality, and conversational dominance. Many researchers have since adopted this model in explaining the different patterns of co-construction (Brooks, 2009; Davis, 2009; Galaczi, 2014; Kley, 2015). Based on this model, there is an association that collaborative pattern of interaction will allow for "better" co-construction. This view is pervasive and has been borrowed to be used in rubrics for communicative competence which also assess other aspects of competence such as grammatical competence (Ockey, 2014; Ockey, Koyama, Setoguchi, & Sun, 2015).

This view and treatment of IC should be problematized for two reasons. First, co-construction is a manner through which participants make sense with and of each other. Jacoby and Ochs (1995) commented that all social actions are co-constructed, so by extension, there could be nothing that is social that is "less" co-constructed, or not co-constructed. Second, without specifying what is being co-constructed, how can we know that being collaborative should be the goal of any testing situation?

It might be a misconception, however, to think that IC is entirely co-constructed. With conversation analysis methodology, Kasper and Ross (2013) pointed out that we can assess individual IC in any joint formation, provided that we look at one's contribution to the ongoing talk in relation to what came before his/her turn, sequence, or even to overall trajectory of activity, and considered how those prior actions constrain or enable his production.

> Participants in interaction are both constrained and enabled by the co-participants' actions. In that sense, interactional competence is distributed between the participants; no-one owns it. But it is equally important to recognize that the coparticipant's prior turn opens up an opportunity space for the current speaker's actions, it does not determine them. Individual participants do show themselves as more or less interactionally competent at particular interactional moments. With these cautions, we proceed. (Kasper & Ross, 2013, p. 11)

To guide our attention to different layers of actions in interaction, language assessment can borrow from CA studies, which shares the interests on the construct of IC. Nguyen (2012a) suggests that we can examine an individual's competence display while paying attention to how the competence is co-constructed by participants in interaction. In a testing or assessment situation, raters can look for evidence to support how, in a given specific set of actions and situations, individuals can show themselves to be more or less interactionally competent, and this could be viewed as the person's interactional competence within that co-construction (Kasper & Ross, 2013). A skilled participant with strong interactional competence is someone who in that situation can make good use of the resources that are available. These resources include linguistic and pragmatic resources to the practices – knowledge of rhetorical scripts, knowledge of lexis and syntactic patterns, knowledge of turn-taking management, knowledge of topical organization and the means to transition from one topic to another (He & Young, 1998).

IC in Language Assessment Through a CA Lens

Kasper and Ross (2013) classified language testing research interests relating to the construct of IC based on two concerns: first, to describe the interactional organization of oral performance assessment with claims to assess learners' second language interactional ability, and second, to define and operationalize IC as part of the target assessment constructs.

This first line of interest consists of a larger body of literature describing interactional features of different types of oral performance tasks. Even though the studies under this strand may not always brand themselves as being directly related to IC, it is important to recognize that these studies have illuminated such rich descriptions of test-related IC required of test takers in different types of elicited performance. This consideration of IC is therefore directly related to the construct validity of any test instruments whenever they are being used to elicit some form of oral interaction production. This is because knowing what actions are being carried out, by what resource and how, is closely linked to the issue of task design and the construct validity of the task, whether it assesses what you intend to assess.

Assessment performance as social practices. With respect to providing interactional descriptions for multiple forms of oral assessment, the oral proficiency interview (OPI) has by far been investigated the most given the long history of its applications when it comes to assessing oral ability in many high stake tests like the ACTFL OPI (Lazaraton, 2002; Tominaga, 2013) and the IELTS speaking test (Seedhouse, 2013; Seedhouse & Egbert, 2006). Contemporary findings have agreed that the language samples elicited through the OPIs underrepresent the construct of conversations in everyday contexts because of their considerable differences in sequential structures, topic organization, preference organization. Also, the pre-specified turn-taking allocation structure between examiners and examinees are vastly different from that of ordinary conversations (He & Young, 1998; Lazaraton, 2002, 2008; Van Lier, 1989;

Young & Milanovic, 1992). Recent research on OPIs acknowledges that OPI are a specific kind of institutional discourse, a tool for eliciting evidence of spoken language 'proficiency,' and an occasion designed for test takers to produce language samples which demonstrate their control over various discourses and grammatical structures (Ross, 2017; Seedhouse & Nakatsuhara, 2018).

On treating the OPI as an institutional practice with the specific goal to elicit proficiency displays, some studies have focused on describing the examiner's practices during OPIs. Research findings revealed the extent to which the examiners influenced the language samples during the OPI interactions to help examiners become more efficient in reliably eliciting ratable speech samples (Kasper, 2006; Kasper & Ross, 2007; Lazaraton, 2002; Seedhouse & Egbert, 2006). For example, in a high stake speaking test like IELTS, examiners are advised not to initiate repairs when the examinees seem to misunderstand their questions or when their responses appear incomprehensible, as the inability to provide relevant answers is already serviceable as evidence for language proficiency placement (Seedhouse & Nakatsuhara, 2018). Other studies have focused on describing the interactional features of test takers which can go into the revisions of task design or the adjustment of its assessment criteria. For example, Tominaga (2013) investigated storytelling sequences, produced as part of ACTFL OPIs, and described how the task afforded opportunities for the test takers to display their Japanese L2 IC. She found that although the test takers' storytelling had observably become more effective at the time of the re-test, their scores did not necessarily reflect that. Her study underscores the importance of rating scales that are sensitive to interactional improvement in learner's social practices.

While the OPI format is recognized for its shortcomings in eliciting the range of interactional abilities needed for everyday interaction, roleplays are seen as a viable eliciting tool

for a wide range of interactive practices since they are not restricted by the interviewer-guided question-answer structure (Grabowski, 2013; Kasper & Youn, 2018; Kormos, 1999; Okada, 2010; Okada & Greer, 2013).

An ethnomethodological (EM) approach which informs CA's take on roleplay has been instrumental in countering the negative associations of roleplay and simulation with inconsequential, inauthentic, and unnatural interaction,. from the empirical focus on how roleplay participants manage their interactions under the parameters of roleplay in the first place (Kasper & Youn, 2018). EM study roleplays as a social activity, and its findings display the order of roleplay interaction from the participants' emic perspectives. Therefore, the question of whether the interaction is unnatural or inconsequential becomes an empirical inquiry into how the participants treat the interaction as such (Huth, 2010). In particular, EM analysis on roleplay and simulations helps elucidate how participants invoke the roles or identities specified in the roleplay set up through employing membership knowledge of the social categories associated with those roles (Watson & Sharrock, 1990). Meanwhile, Francis's (1989) analysis of simulated interaction reminds us that regardless of whether one participates in a roleplay or simply interacts as oneself, the contingency of interaction and its locally oriented character still remain the same (p. 54).

In ACTFL OPI roleplays, where the interviewer and interviewee would make a transition from the OPI portion to the roleplay portion as part of the test progressivity, Okada (2010) reported that to competently complete the roleplay, test candidates have to display their knowledge of the appropriate sequence organization of the action required by the roleplay which can be drawn directly from ordinary conversation. To illustrate this point, a transcript from Okada (2010) is reproduced below in Excerpt 2.3.

Excerpt 2.3 *Okada (2010, p. 1660); "Cleaning Shop" (I: Interviewer, C: Candidate)*

```
35     I:     yes ma'am, may I help you.
36     C:     yes uh my jacket is covered with mu:d uh:::
37            mm (1.9) can you clean it for me?
```

In this situation, the candidate assumed the role of a customer while the interviewer was playing the role of a shopkeeper. In producing her response in line 36-37, the candidate displayed an understanding of the previous action in line 35, and the social activity which normatively takes place at a cleaning shop. The candidate also displayed her IC in producing her request in a sequentially appropriate manner, with an account in line 36 that functioned as pre-request.

Interestingly, Okada's (2010) observation of the examiner's performance during the roleplay revealed a still-asymmetric turn-taking allocation and rights to topic nomination. Given that the roleplay is part of the OPI test, the interviewer remained the gatekeeper of when to move on the next topic and what topics got blocked or ratified in the roleplay conversation. Okada and Greer (2013) reported that the interviewer would pursue a task-relevant response by employing multiple questions or withhold their uptake of the candidate's actions which did not follow the activity required by the task, resulting in extensive delays in which the candidates could then self-repair to correct their action.

To sum up what we learned so far about roleplays, the common requirement which the participants need seems to be the membership knowledge of what playing the roles would generically entail. This includes the knowledge of the social activities which are categorically bounded to a certain social group. When participants are asked to roleplay as "themselves" (i.e., Walters, 2007) or any character that they "know" how to play, their interactions are therefore authentic in the sense that the participants bring personal biography, epistemic resources, and membership knowledge to display and enact a social activity in the roleplay.

For these reasons, roleplay is seen as an appropriate test format particularly for assessing pragmatic competence in interaction (Grabowski, 2013; Kasper & Youn, 2018; Youn, 2015), and by extension for assessing social actions and courses of actions, provided that the roleplay, "as an interactionally constituted activity, is seen as affording the necessary infrastructure for examining how test takers produce and understand social action-in-interaction through turns and sequences" (Kasper & Youn, 2018, p. 4)

IC as target construct of assessment. The distinction between IC in assessment summarized earlier and the IC assessment addressed in this section is that while the former's goal is seeking to describe, the latter's goal is to measure. Therefore, this section discusses instrumental designs and assessment benchmarks that have been established for assessing L2 IC operationalized under the EMCA approach.

Existing findings from IC development on L2 data suggest a positive correlating trend between language proficiency and IC as operationalized in those studies (i.e., Al-Gahtani & Roever, 2018; Hellermann, 2008); however, to assume that IC will always correlate with language proficiency would be a mistake. Studies which have investigated IC development in professional discourse, such as Nguyen (2012b), showed that even highly proficient speakers fine-tuned their interactive practices while adapting into a new institutional activity that was not familiar to them. One key consideration that we should keep in mind is that IC is activity specific. While making predictions about future performance within the confines of similar actions or courses of actions might be allowable, generalizing IC measured in one activity type onto another action will be highly problematic.

In the earlier form of incorporating social constructs into the assessment of 'language use' under the framework of communicative competence (Bachman, 1990; Canale & Swain, 1980), much work centered on assessing pragmatic competence (Brown, J. D., 2001; Brown, J.

D. & Ahn, 2011; Grabowski, 2013; Hudson, 2001; Hudson et al., 1992, 1995; Roever, 2011; Yamashita, 2008; Youn, 2013), which was defined as cross-cultural pragmatic ability in three speech acts: request, apology, and refusal. The second wave of the pragmatic assessment derived from this framework has since then been extended to incorporate pragmatic competence in interaction (Ross & Kasper, 2013; Youn, 2013, 2015), which taps more into the construct of IC.

In assessing pragmatic competence in interaction, Youn (2013, 2015) designed two open roleplay tasks, one in which the test takers interacted with an interlocutor and another in which they interacted with a partner, to elicit their performance on doing requests and agreeing on a meeting time. The rating criteria of her study included the following: *content delivery*, *language use*, *sensitivity to the situation*, *engage with interaction*, and *turn organization*. Youn (2013, 2015) operationalized in detail the target interactional accomplishments for each criterion which is particularly sensitive to the consideration of IC. In the two criteria which directly address interactional aspects, *engage with interaction* and *turn organization*, Youn included the following considerations in her rubric's descriptors. Under the criterion *engage with interaction*, raters were directed to check whether the students designed their turns in a way that cohesively responded to the prior turns. Under *turn organization*, students' performances were checked for adjacency pairs completions in a time-sensitive manner (Youn, 2015). Interestingly, her quantitative findings through multifaced Rasch analysis or FACETS (Linacre, 1998b) indicated that these IC related criteria were the two easiest criteria in her study, pointing to the possibility that the L2 speakers might have developed IC well before they mastered other aspects of L2 competence.

As for an assessment of IC on its own, there has not been any assessment framework which has operationalized the construct consistently or systematically. On the reasons of this difficulty, Ross (2018) cited IC's multicomponential nature which as a consequence makes it

hard to operationalize and the fact that observing IC in real-time can be too subtle to warrant IC as an independent criterion.

However, there is currently a prevailing push in this direction given its potential contributions to assess the construct which has been lacking in oral proficiency assessment. In a recent special issue in the journal *Language Testing*, the authors explored the construct of IC and its related issues including the relationship between IC and proficiency level, listener response as a distinct feature in the IC construct, assessment formats and tasks that can elicit IC display, and the specific criteria that are incorporated into a rating scale for IC assessment (Plough, 2018). Notable findings which pointed to the direction of assessing situated and action-specific IC can be drawn from Ross's (2018) study, which revealed how listener responses as part of IC are dependent upon the task, and Kim's (2018) study, which analyzed communication for specific purposes and reported that in an interaction between L2 air traffic controllers, the profession-specific competence is perceived as more consequential than language proficiency.

While it is challenging to operationalize IC for measurement purposes, there seem to be an agreement that the field cannot "dispose of the difficulties by simply defining them out of existence" (McNamara, 1996, p. 83). Given the current research's attempt to propose and validate a framework for assessing IC in this chosen group roleplay interaction, the next and final section of this chapter provides a brief overview on the concept of validity in language testing and assessment.

Validity and Validation in Language Assessment

The concept of validity is one of the key topics of studies in the field of educational measurement which has provided the basis of our understanding of valid observation in many disciplines aiming to measure unobservable cognitive psychological traits like "knowledge" or

"mastery" of any skills which have been learned or acquired. It should be noted that outside the discipline of language assessment, other disciplines within the measurement community debates validity and validation largely without referencing any specific content. Language testing is unique in that it has a content domain as the focus of its research, and this makes a difference to the kinds of challenges that we face in approaching validity from a language testing perspective (Fulcher, 2015, p. 107).

With that being said, language testing has borrowed considerably from the field of educational measurement (Chapelle, 1999). To date, there are three major waves of development for the consideration of validity which have been influential in the field of language testing. These waves were spearheaded by the work of Cronbach and Meehl (1955), Messick (1989a, 1995, 1996), and Kane (2006).

In its earlier theory, validity is defined as the degree to which a test instrument is measuring the construct it claims to measure. According to the validation consideration proposed by Cronbach and Meehl (1955), validity is divided into three categories: criterion-oriented validity, content validity, and construct validity. To gather evidence in support of a test's criterion-oriented validity, testers would take interest in checking the empirical relationship of its scores with a criterion. The criteria could be one or more reputable tests that measured a similar construct (concurrent validity) or an external criterion in which the test wants to make predictions (predictive validity). The stronger the correlations between the test and the criterion, the stronger the case our evidence for validity is. Content validity is confined by the extent to which the test content is representative of the domain knowledge we wish to test. Content experts provide their judgments on test content representativeness regarding the skills, the context, or the scope of knowledge being sampled for the test. Construct validity concerns with the extent to which the operationalization of the construct aligns with the theory underlying the construct. It

also requires an array of evidence which could support the interpretation of the scores obtained from the tests.

Following Loevinger's (1957) argument that based on this model, criterion and content validities can be subsumed under construct validity, Messick (1989a) proposed a unified model combining the three types of validity under one consideration over construct validity. Under this unified framework, he added that validity consideration must also have an appraisal of the social consequences of the test implementation. This component expands the validity argument to cover not only the internal ecology of the test but also the social consequences of the test. Messick's framework has been highly influential as it refocuses the concerns of test validity away from merely the property of a test itself to the validity of test scores and the meaning that we attach to them. It emphasizes that a test does not happen in a vacuum and to consider the property of a test without paying attention to what it is used for cannot provide a complete picture of its validity.

By unifying the validity consideration under one umbrella of construct validity, Messick has made it more evident that validity is "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p. 13). This approach to validity has been very prominent in language testing research (McNamara, 2006) as it provides a comprehensive process for test validation. Also, another huge impact from Messick's unified theory of validity is that the social consequence component in his model has provided the grounds for discussions of societal values, test impact, ethics, and washback to be included in the validity conversation itself.

Kane's (2006) interpretive argumentative approach to validity derived the basic tenets from Messick's argument-based model. The distinction between Kane's and Messick's models

lies in their emphasis on the underlying "construct." Fulcher (2015) noted of Kane's general distrust of the naming of the abstract concept based on a set of observations and then assuming that the named "thing" exists. Alternatively, Kane proposed a move away from directing our attention to construct validation towards giving more primacy to observable attributes derived from target domains and task descriptions (Kane, 2013, pp. 21-24). For language testing, Chapelle, Enright, and Jamieson (2010) stated that Kane's organizing concept of an "interpretive argument," which does not rely on a construct, has proved to be useful because the link is to be established directly from observation to claim without the need for constructs at all.

Kane's approach to validation sees that what is most necessary to the validation process are the ties between observational attributes and a particular scoring decision. These ties can be established from something as simple as a chain of logical reasoning in the form of 'if-then' structure. Then, the work of validation study would be the establishing of a rationale for such links. Kane's version of validation provides an explicit, yet contingent, approach which lies on the strength of the argument you make about the validity of the scores' meaning, and this could be driven by both the theory or practicality of test use. This emphasizes the importance of due process in assembling a validity argument. To Kane, an argument is made to support a claim, backed up by evidence (Kane, 2016). The counter-argument must also be considered before the interpretation of test scores' validity is then evaluated.

From the evolution of these approaches to validity and the process of validation, we can see that the early theories of validity fit well with the discrete type of language tests, like those composed of multiple-choice items or gap-filled items, which dominated the landscape of language test practices at the time. With the communicative movement in language teaching and testing which started around the 1980s, an approach to validity and validation allows for less

concentration over psychometric properties of the test and more to compiling pieces of evidence in support of the kinds of claim we are making about the scores (Kane, 2017).

With Messick's and Kane's interpretive argumentative frameworks, language tests have a tool to study validity while still exploring a more communicative, situated, and socially ingrained construct such as IC. In this early stage of expanding our understanding of the construct, we are reminded of Messick's recommendation that the process of validation does not end at any one point. As long as the scores still hold a meaning, across different contexts, times, or populations, validity is a perennial empirical question, an evolving property which requires a continuing process of validation (Messick, 1995).

CHAPTER 3

RESEARCH FRAMEWORK

Given the juxtaposition of the topic of interactional competence (IC) and its application in the field of language assessment, this study found itself in the nexus of different methodological families. As we have seen the previous chapter, the construct of IC has almost exclusively developed as a subsidiary field of conversation analysis, which fundamentally operates from an emic perspective or the participants' point of view. On the other hand, the whole enterprise of language testing is operating on an etic perspective. For these reasons, Mixed Methods Research (MMR) has been chosen to provide the macro analytical framework for this study, given its pragmatist stance and its flexibility, both of which allow for a pursuit of systematic answers to take priority over methodological boundaries between different research methods or paradigms.

This chapter presents a literature review on topics relating to the methodological underpinnings of this study. First of all, it will lay down key issues of MMR as our macro framework and discuss key studies in the field of language testing which utilized MMR as its research method. Second, moving closer to methodological concerns surrounding the construction of a new rubric, the second section of this chapter will give a comprehensive summary of rubric and rating scale construction, especially ones designed for assessing task-based language performances. Finally, the last section will cover the issue of applying conversation analysis for a non-naturally occurring talk (especially roleplay data) which is not a common object in CA, but quite common in language testing data and definitely applicable for this study as well. The conclusion of this chapter will address any potential challenges and controversies which could arise from our MMR study design.

Mixed Methods Research

Mixed methods research or MMR, often referred to as the third major research paradigm (Johnson, Onwuegbuzie, & Turner, 2007), can be conceptualized most simplistically as a research method which combines two or more qualitative (QUAL) and quantitative (QUAN) methods within a line of research inquiry. While MMR is recognized for its methodological mixing, Jang, Wagner, and Park (2014) emphasized that its scope extends beyond method mixing and involves the entire process of inquiry. This echoes what Greene (2008) noted, that MMR constituted a distinct approach toward social science inquiry in itself, and that merely mixing QUAL and QUAN research methods in one study does not make the study MMR.

This middle ground position and the pragmatist logic that MMR adopts provide a hint of previous struggles that shaped the principles and underlying philosophy of this method, often referred to as the "paradigm wars" (Gage, 1989). On one side, there were the positivist/ objectivists who were branded under quantitative methods, and on the other, there were the constructivist / interpretivists and others who pushed for the legitimacy of qualitative methods. Critics of QUAN methods pointed to the fundamental difference between human social sciences disciplines and the discipline of natural sciences which the positivists were trying to emulate. In reality, human beings do not always behave in a stable manner like rocks or water; their behaviors could easily be affected by the environment, and the interaction is rather unsystematic.

Given this natural complexity inherent in the phenomena that we wish to study, QUAN's tendency to reduce their observation to just numbers can be overly simplistic and reductive. However, using numbers and statistics, when done properly, can provide a defensible and authoritative account of the phenomena we wish to study. QUAN methods can supply breadth to the description of the behavior of large groups and provide an explanation in probability terms (Brown, J. D., 2014). Borrowing the tools for quality assurance from the field of natural sciences

means that QUAN methods also have many sophisticated ways to guarantee the quality of QUAN empirical research.

Given these supposed strengths of QUAN methods, the paradigm wars reminded us that the methods also have its limitations. Dörnyei (2007) commented that numbers are powerful, yet they are powerless in themselves as they are faceless and meaningless unless we have given these numbers the precise definitions to back up the validity of their usage (pp. 32-33). Ultimately, the paradigm wars served as a reality check for researchers to note that the method has to serve their research questions, not the other way around. And while proponents of MMR who adopt the pragmatist or methodological eclecticism stance do argue that it is acceptable to combine methods by choosing what they believe to be the best tools for answering their research questions, it is also important to note that the best method for a given study may be purely QUAL or QUAN method rather than mixed (Teddlie & Tashakkori, 2010).

Given its pragmatic approach (Johnson et al., 2007), or what Teddlie and Tashakkori (2010) called a methodological eclecticism approach to its scientific inquiry, MMR sees research techniques in QUAL and QUAN as resources from which researchers can select what is most appropriate for the purpose of more thoroughly investigating the phenomenon of their interest. Recognizing the strengths and limitations of both QUAL and QUAN, MMR attempted to respect both research traditions while trying to strike for a workable middle ground in service of the research problems at hand.

Teddlie and Tashakkori (2012) summarized the history of the development of MMR as a research paradigm that we have discussed before into the following nine characteristics: (a) methodological eclecticism, (b) paradigm pluralism, (c) emphasis on diversity at all levels of research enterprise, (d) emphasis on continua instead of dichotomies, (e) iterative and cyclical approach to research, (f) centrality of research questions, (g) explicit discussion of research

design and analytical process, (h) tendency towards a balanced middle-ground standpoint rather than polarity, and (i) reliance on visual representations.

One issue which has been sidestepped in this paper so far is the concerns over paradigm incompatibilities when combining QUAL and QUAN methods. To a true MMR believer, the principle of methodological eclecticism would rule that incompatibilities do not exist. Many mixed methods researchers, however, would say that this is not always the case. Dörnyei (2007) said that this is not a matter of black and white, as it depends on the different conditions surrounding your research topics, contexts, and ultimately your research questions.

MMR Typologies and Designs

Given the fact that MMR studies tend to be situated and respond directly to research questions rather than following any pre-set formula, there are numerous ways that a study of language assessment can make use of MMR designs beyond any typologies to fully account for the varieties (Maxwell & Loomis, 2003). However, it is still useful to discuss the common framework which categorizes MMR in terms of its purposes and design characteristics (Greene et al., 1989) for the sake of uniformity of references terms when discussing the design of any MMR studies.

The typological framework commonly used for categorizing MMR studies has been the one proposed by Greene et al. (1989) in which the authors delineated methodological purposes and methodological considerations of MMR into five families: triangulation, complementarity, development, initiation, and expansion. A summary of the aforementioned purposes and design considerations are summarized in table 3.1 below.

Table 3.1
Summary of MMR Methodological Purposes and Design Considerations from Greene et al. (1989)

Methodological purposes		**Methodological design considerations**	
Characteristics	Purposes	Characteristics	Considerations
Triangulation	To confirm or cross-validate findings from different sources.	**Phenomena**	Do QUAL and QUAN gather data from similar or different phenomenon?
Complementarity	To elaborate, enhance, or illustrate the results from the other.	**Paradigms**	Do QUAL and QUAN come from the same paradigm?
Development	To use results from one to help inform the development of the next methods.	**Status**	Do the QUAL and QUAN have the same level of importance?
Initiation	To uncover potential paradox and contradictions underlying research problems.	**Implementation: interaction**	Are the QUAL and QUAN implemented interactively or independently?
Expansion	To expand the scope, breadth, and range of the study to include different components.	**Implementation: timing**	Are the QUAL and QUAN implemented concurrently or sequentially?

To recognize the distinctions among different research designs, Greene et al. (1989) proposed five primary considerations in which MMR researchers should determine as they try to configure the design of their MMR studies: first, the phenomenon or phenomena being investigated under different research methods; second, the research paradigms from which each of the methods are derived; third, the relative weight of importance given to each research method; fourth, the degree in which different methods will interact throughout the process of implementation; lastly, the timing of each methods whether they would be implemented sequentially or concurrently.

With these influential typologies outlined by Greene et al. (1989), we are able to capture many different ways that MMR can be designed and used. Personally, I think that this framework can cover quite an extensive typology of MMR. However, as MMR will continue to be used to address complex issues, and diversity of designs and interpretations will still be paramount to MMR, this typology shall only be used as a guidance, not a restriction.

MMR for Language Testing and Assessment Research

The field of language testing and assessment has traditionally been dominated by quantitative positivistic research paradigm. Jang et al. (2014) attributed the recent shift away from this longstanding trend within the past two decades to two factors: (a) the expansion of theoretical definitions of language competence (Bachman, 1990; Canale & Swain, 1980; Hymes, 1972), and (b) the expanding framework of test and assessment validation (Messick, 1989a, 1995, 1996). Researchers on language testing and assessment are increasingly turning to MMR in order to understand the complexities of language acquisition, interaction among language users and their impact on language testing and assessment (Jang et al., 2014). MMR has become instrumental in investigating validity claims beyond the three classical facets of construct, content, and criterion validity, allowing researchers to consider social aspects of test use such as its social and political influences, and their impact on learning and teaching.

In second language assessment, MMR has a great potential to be used in research relating to different facets of assessment validation (e.g., Baker, 2010; Lee, Y. & Greene, 2007; Youn, 2015), or second language communicative performance generated for testing and assessment purposes (eg. Nakatsuhara, 2013) as the field has become more and more aware of the complex social issues relating to language testing and assessment (McNamara & Roever, 2006).

Challenges in Mixing Conversation Analysis with Quantitative Methods

Because of CA's firm root in ethnomethodology (EM), which is radically emic in their stance towards explaining social organizations, it generally shows apathy towards other methodologies with externally motivated theoretical preconceptions such as hypothesis testing, an application of coding schemes, or any generalization beyond the situated context at hand (Garfinkel, 1967). However, as the field of CA has grown, contemporary researchers have

argued that it would be useful to see how the existing CA research might be applicable outside of CA itself (Antaki, 2011). (For recent studies which employed CA in sequential mixed methods research on institutional interaction, see de Ruiter & Albert, 2017; Kendrick & Holler, 2017).

At this point in time, applying conversation analytical method in quantitatively oriented studies is not something unprecedented. Many studies in conversation analysis in the field of second language acquisition (CA-SLA) have recognized that in order to answer research questions regarding the development of L2 learners, descriptive findings from conversation analysis (CA) alone can be insufficient. Tracing developmental changes in L2 use across time and proficiency levels is only possible when there are recognizable objects of learning for analytic comparison (Lee, Y.-A. & Hellermann, 2014). To this end, CA-SLA research pursuing longitudinal or cross-sectional agendas has worked on ways to establish analytic parameters for identifying and tracing social phenomena and how learners differentiate their practices across time. The challenge in doing so is in establishing that any differences observed in learners' participation methods can be attributed to the change of their language competencies, or if they are simply an artifact of different situations and circumstances facing the learners at the time.

Koschmann (2013) suggested that any method used for classification in EM must address "locational concerns in microanalytic terms" (p. 241). In resolving the methodological challenges regarding comparability and quantifiability of talk-in-interaction, studies refer to Schegloff (1993), who recommended that analyses seeking to compare or quantify interactional practices or actions require (a) a defensible specification of sequential environments in which that practice or action can take place, (b) an understanding of possible alternative practices and actions which can occur in that same spot or sequential environments, and (c) a defensible specification of the domain of activity from which the observations are drawn and from which the inferences are made. In other words, comparative CA studies need to identify stable

sequential environments where an interactional practice or action can regularly be identified, and the absence of such action is treated as interactionally consequential.

The issues of comparability and traceability facing longitudinal as well as cross-sectional research on interaction are directly relevant to the development of an assessing instrument which seeks to categorize learners based on their displayed interactional accomplishments. It is believed that the methodology used in identifying interactional phenomena in longitudinal research can be applied to the development of an instrument for IC assessment purposes.

One of the goals of the current study is to explore this methodological application in the development of a rating scale to assess IC in a peer group roleplay activity. In developing such the test, it is vital to first identify the social actions or courses of action which can allow for sustainable comparison across the population of learners we need to assess. The second step, which is unique to assessment-motivated research agenda, is that there be an institutional need to rank and ascribe values onto different participation methods within the identified actions. Depending on the granularity of the interactional order which the test instrument is targeting, albeit at the level of turns, sequences, or a course of action, existing findings from CA can serve as valuable resources that test analysts can reference in guiding the test's judgments in making arguments regarding the validity of the test being developed.

Performance Assessment

Based on the definition of IC which has been discussed in the previous chapter, one significant implication for language testers is simply that IC cannot be separated from performance (Roever & Kasper, 2018) as it is only observable in interaction. This brings to the fore other issues which have arisen with performance assessment including the validity of

observation or elicitation, assessment criteria, and raters' behaviors. This section summarizes the discussion concerning performance assessment and its development and validation.

Performance assessment takes direct observations as the way into evaluating test takers' ability on a given construct of interest. Its direct approach to language assessment is often advocated for its authenticity given that it can be more contextualized to approximate 'real-life' tasks and be judged on authentic standards. Teachers have argued for the benefits of using task-based or performance assessment, especially when teaching for specific purposes (LSP), saying that it enables complex, integrative demands of language use, which is more realistic to how language is used in real life (Linn, Baker, & Dunbar, 1991).

Inherent to any use of performance assessment is that it has to be subjectively scored by human raters. Having raters as an intermediary agent between the examinees' ability and the observation of the construct in question poses a double threat to test validity as both the task and raters can now introduce a degree of construct irrelevance in the variation of test scores which could weaken the validity of inferences we wish to make about the examinees. The specification of assessment criteria is therefore paramount in influencing the validity of test results since the criteria offer a model of construct representation which has a direct relationship to the operationalization of the target construct (Norris, 2001). However, the identification of relevant and valid criteria has also been a challenge for language testers (for examples, see Elder et al., 2012; Knoch et al., 2015). Examinees' performance requires human raters to determine the quality of their performance against some kind of rating scale. IC is unique in that it can only be observable in an unfolding interaction, as hearable or visible social actions, action trajectories, and practices (Roever & Kasper, 2018, p. 333). Therefore, it is a competence which cannot be elicited through any indirect measures, except in the tasks which include some element of talk. For this reason, the reality facing practitioners who wish to incorporate IC construct as part of

their assessment criteria is that they must figure out how to handle issues relating to raters and the criteria on which they want their students' demonstration of IC to be judged.

Raters

It has long been recognized that raters bring a considerable about of variability to the raw scores. This variability derived from raters' characteristics is seen as undesirable, and traditional approaches to improving rating processes have always looked to eliminate raters' influence through training and accreditation process (McNamara, 1996). Studies which have investigated the effect of rater training have shown that successful rater training can make raters more self-consistent (Eckes, 2011; Weigle, 1998), and a higher degree of agreement among raters can be achieved (Davis, 2015), resulting in a more homogenous rater performance. However, different degrees of rater severity seem to persist regardless of training or experience (Lim, 2011; Lumley, 2005; Lumley & McNamara, 1995).

While language testing researchers still have yet to have a definitive answer to questions relating to the characteristics and behaviors of raters and the interaction between the raters and rating scales, from a test developer's standpoint, score variability from the rater component has become something that is manageable thanks to the multi-faceted Rasch measurement model developed by Linacre (1989, 2006). The FACETS computer program provides an estimation of examinees' ability in performing the test task, taking into account the characteristics of other facets, i.e., the raters' behaviors, which may have influenced the raw scores (McNamara, 1996). Researchers can then follow up on cases, such as any particular examinee's performance, which demonstrably do not fit the model to find out what might have contributed to such result (Bond & Fox, 2015; Linacre, 2002).

With this understanding, the goals of rater training, therefore, is not to eliminate variability among raters, but to lead raters to an understanding and application of the scoring criteria that accurately reflect the construct that the test is intended to measure. Training sessions are to help raters focus their attention on the elements of performance targeted by the assessment (Fulcher, 2003), as well as to standardize the raters' perceptions to minimize extreme judgments or ratings (Weigle, 1994).

Rubrics

Decisions which test designers made about the criteria on which performance will be judged are often made explicit in the form of a rubric. For classroom assessment, a rubric is a tool that any language teachers can use for scoring students' language abilities and, perhaps more importantly, for giving them feedback on their progress in learning those language abilities (Brown, J. D., 2012, p. 1). For larger scale assessment purposes, rubrics can help enhance the reliability of judge-mediated ratings and limit possible rater biases (Slater, 1980). Altogether, it is undeniable the crucial roles that rubrics play in warranting validity claims of performance assessment design. They are used by raters to guide their rating process, and they are also the means through which the scores are reported and interpreted (McNamara, 1996).

Fulcher, Davidson, and Kemp (2011) have described two approaches available for designing a rubric for language assessment purposes: the measurement-driven approach, and the performance data-driven approach. With a measurement-driven approach, the criteria included in the rubric reflect the theoretical communicative construct. The performance data-driven approach, on the other hand, treats the data as the primary source for rubric construction as it seeks to describe performance data in detail before organizing the performance into levels which will then be used as descriptors for the rating scale. Fulcher et al. (2011) argue that the data-

driven approach has an advantage over the measurement-driven approach because first, it generates richer and better-fitted descriptors for the contexts and other performance conditions which can potentially increase the reliability of the rating instrument; also, it allows for creating a diagnostic profile which could provide users with more relevant feedback for language learners.

For assessing IC, the data-driven approach to rubric construction is a crucial step for rubric construction as only through the actual performance can the interactional practices be identified, the interactional goals be established, and the relevant actions be selected to represent the IC construct we wish to assess. To this end, as discussed in more detail previously, applied conversation analysis for can provide a methodological framework to identify recurring interactional phenomena in performance data in the microanalytical detail necessary to identify the IC necessary for effective interaction for that particular activity (Schegloff, 2006) that we can set out to assess.

Scale Development and Validation

Rubric constructions within the data-driven approach begin with a rich description of how students performed the tasks, and the descriptors stating the skills or behaviors required to successfully perform for each rating criteria are derived from such qualitative analysis (Fulcher et al., 2011). The process of validating the scoring levels in constructing a well-functioning scale, however, has rarely been reported.

Decisions involving the number of levels to include in a scale require a balancing act between theoretical and practical considerations. On the one hand, the language-related theories which inform the definition of the target constructs can provide a guiding framework to inform the test developer on the decision (Lantolf & Frawley, 1985). However, it has been reported that

the theoretical account alone can lead to a poor fit between the rating scale and the test data (Knoch, 2010; Mendoza & Knoch, 2018). Weigle (2002) cautioned that having more levels can increase a risk towards scoring reliability as there are limits to the number of distinctions raters can discern. To validate the use of any rating scale, therefore, requires an understanding of how raters interact with the scale in a continuous process of operating the scale to gain insights into how the scale is functioning.

This requires the help from a quantitative analysis. The multifaceted Rasch measurement (MFRM) model has been utilized to provide such diagnostic analysis in L2 performance assessment studies (Knoch & Chapelle, 2017; McNamara, 1996; Mendoza & Knoch, 2018) as the MFRM model analyzes the scores as a function of the interaction of student abilities, task difficulty, as well as rater severity (Bond & Fox, 2015). Through this model, we can see that a well-functioning scale should be able to capture the whole range of student abilities on the given task and allow for a clear differentiation between each score step that matches the differentiation of abilities in each score level (Myford & Wolfe, 2004).

Within an argument-based framework for assessment validation (Kane, 1992, 2006, 2016), constructing a validity argument for assessing language performances requires an explicit specification on all the inferences connecting the test takers' performance to how the scores are used and interpreted. Rating scales provide the link between the scores and the test construct, and they are also the lens through which test-takers' performances are judged, and the test results are interpreted (Knoch & Chapelle, 2017). Throughout the process of validating and developing a rating scale, mixed-method research (MMR) provides a necessary framework and a unifying philosophy to tackle a complex task of developing and validating an assessment of interactional competence as its construct.

CHAPTER 4

METHOD

The mixed methods research (MMR) approach (Brown, J. D., 2014; Dörnyei, 2007) is adopted as a macro framework for the data collection and analysis of this study. A combination of qualitative and quantitative methods will be used at different stages for this investigation corresponding to the research questions at different stages of the study. Justifications for why methods are used at different stages are discussed along with the design to argue for the validity or trustworthiness of using mixed methods research. The following sections describe the assessment context, the research questions, the procedures of data collection, and analysis of the data from the qualitative and quantitative phases of the study.

Assessment Contexts and Assessment Task

This section describes the contexts of the chosen performance data from a peer group roleplay assessment task. Set in the context of English as a foreign language curriculum for undergraduate level at a university in central Thailand, this study is aimed at exploring individual and co-contributions within EFL IC in group oral tasks using role-play simulation in order to inspect, expose, improve, and validate the standards of IC that are imposed on and oriented to the students in response to a socializing in semiprofessional environment. The course in which this *socializing task* is situated is a 16-week-long English for engineering course titled *Communication and Presentation Skills*. This course is a required undergraduate course offers twice a year to students in the faculty of engineering from all of its majors. Most students take this course during their second year, but some may take the course in their third year depending on their majors. The course objectives are targeted at helping improve students' oral communication skills, covering a number of communicative activities including socializing talk,

problem-solving discussion during meetings, job interviews, and presentations on engineering related topics. The class meets three hours per week, during which class of instruction and assessments take place. The course content is divided into four units corresponding to the curriculum goals: socializing, meeting discussion, job interviews, and presentations.

The primary data for this study come from the first unit on socializing talk. The summative test at the end of the unit is designed to elicit group interaction, specifically everyday small talk for working and networking purposes in an engineering context. The activity lends itself well to an examination of interactional competence for a specific institutional purpose.

The assessment task. As specified in the syllabus, the students were to create a persona and play the role of a representative of a chosen company as they attended an international trade show in Sydney, Australia. The scenario required that they were meeting at a pre-conference reception and that none of them was from Thailand. In preparing their own roles for the role-play, each student was required to choose a name, the company they were going to represent, their job position, and their responsibilities in their respective companies. They also had to research other relevant information such as their hometown, education, etc., as part of their preparation for the roleplay assessment task (see Figure 4.1 for the set-up sheet.)

On the day of the test, five to six examinees were randomly assigned to form a group that took the test together. Each group had 15 minutes for their preparation before performing the roleplay in front of an examiner for 12 – 15 minutes. During that time, the students engaged in and elaborated on the activities specified on the set-up instructions: they conducted small talk, introduced themselves, talked about their work and the company they represented, and exchanged their name cards as part of making new contacts. The original assessment rubric was available to the students and was used by their teachers. It divided the scores into group scores and individual scores. The group score was graded on a 1-5 scale for "group collaboration." The

individual score was composed of three criteria, content, language and pronunciation, and delivery, each graded on a 1-5 scale, totaling 15 points. (See Figure 4.2 for the original rubric.)

Assessment 1 (10%)

Role-play: A conference welcoming party

Setting: The Products of the Year Conference and Expo is an event being held at the moment in Sydney, Australia. As this event deals with a wide range of businesses, it attracts thousands of participants from all over the world. None of the conference participants is from Thailand.

Assessment Task

You will be part of a group of 5 people who are meeting for the first time. You all work for different companies and are meeting **at the conference welcoming party**. You will introduce yourselves to each other and make small talk for 10-15 minutes. All group members should make use of the appropriate communication strategies, language functions and expressions covered in the first unit.

Preparing for the Assessment:

You will need to research a company and a position that you think will be interesting. You should find out key details about your company as well as the job responsibilities that your position would hold (You may make use of the "Company Profile & Job Description" on page 34 as guideline.) in order to have enough material to fill the 15-minute assessment. You should also research appropriate small talk topics. You won't know which people will be in your group until 15 minutes before your group is assessed. You will not be allowed to use a script.

On the Day of the Assessment:

At the beginning of the assessment, the members of the first group will be chosen. They will be given 15 minutes to practice together. As the first group is called in to perform, the members of the second group will be chosen, and they will prepare while the first group is being assessed. This will continue until every group has performed. During the role play, each student needs to participate actively for the entire conversation.

Figure 4.1 *Set-up sheet for the socializing task*

Group Scores (5)

Scores	5	4	3	2	1
Group collaboration (5)	- members contribute in **equal proportion** - **exceptionally well-prepared** and **professionally helped each other** out during the role-play - 1 min +/- the allotted time	- members contribute in **slightly unequal** proportion - **well-prepared** and **helped each other out nicely** during the role-play - **2** mins +/- the allotted time	- **2-3 members dominate** the role-play - **prepared; not so smooth** when helping each other out during the role-play, **in chunks, a few snags** - **3** mins +/- the allotted time	- **1 or 2 member (s) dominate(s)** the role-play - **individually prepared**, seems **fragmented, little collaboration** during the role-play - **4** mins +/- the allotted time	- **1 member dominates** the role-play - **individually prepared**, seems **very fragmented, no collaboration** - **5** mins or more +/- the allotted time

Individual Scores (15)

Scores	5	4	3	2	1
Content (5) (Talking about personal information, jobs and responsibilities; Talking about your company (accurate information), Business card)	- covers all key parts with substantial details - Accurate information	- covers all key parts with adequate details	- covers some points with appropriate details	- covers some points with explanation	- covers some points with some explanation - inaccurate information
Language & Pronunciation (5)	- **full control** of sentence structure - **no major errors, very few minor errors** - very clear & easy to understand	- **adequate control** of sentence structure - **occasional major and minor errors** - easy to understand	- **some major and minor errors** that are **sometimes distracting but do not interfere** with meaning - moderately difficult to understand	- **noticeable major and minor errors** that are **very distracting** - somewhat difficult to understand by sympathetic listeners	- **many noticeable major and minor errors** that **severely interfere with meaning** -difficult to understand by even sympathetic listeners
Delivery (fluency, performance) **(5)**	- **very fluent and natural** - **no script at all** (notes are okay) - may repeat or stumble **a few times**	- **fluent & natural** - **glances** at note **several times** - may repeat or stumble **several times, but not major**	- **moderately** fluent but **sounds memorized**; a few **pauses here and there but not distracting** - glances at notes often - **no major** breakdowns	- **somewhat** fluent; **some** long pauses - relies on notes from time to time - **a few major** breakdowns	- **not fluent; lots of** long pauses - **heavily** relies on notes - **some major** breakdowns

Figure 4.2 *The original rubric used in-course*

Research Questions

Through reviewing the literature on interactional competence (IC), interactional competence development, and the validation framework for performance assessment, it has become quite clear that there is a ripen possibility to construct and validate an assessment instrument that targets IC through merging applied conversation analysis and the data-driven approach to rubric construction for performance assessment. To this end, the current research proposes an activity-based framework in operationalizing IC as the assessment construct in a peer group roleplay performance data. Through Kane's (2006) argument-based approach to assessment validation, this study is aiming to explore empirical evidence in supporting the claims that the proposed rubric and rating scale are defendable in providing a valid evaluation of IC. The current study's research questions are listed below.

1. From the students' roleplay performance data, what are the constitutive interactional phenomena in the form of actions or courses of action which can be established as the targets for comparing IC across the dataset?
2. What are students' methods, in varying degrees of success, in accomplishing the actions or courses of action identified as the targeted interactional phenomena?
3. How can the rich description of students' task performance inform the data-driven construction of an IC assessment rubric?
4. Given the proposed rubric for assessing IC in this roleplay task, how reliable is the rating process in applying the scale to rate the students' performances?
5. Given the proposed rubric for assessing IC in this roleplay task, are there any detectable biases in how the raters apply the scale to rate the students' performances?

6. Through mixed methods research, to what extent can the current study argue for the validity of the proposed rubric and rating scale for assessing IC in this context? How do the findings from mixed methods help to strengthen the validity argument?

MMR Study Design

For our current study, we adopted the developmental sequential mixed methods design. Developmental design always implements sequential timing in collecting multiple data for analysis in which the result of one type of analysis is then used to inform the development of the next method (Greene et al., 1989). For rating scales used in domain-specific tests, the use of performance data-driven approach to scale construction is frequently encouraged. Students′ performance data are treated as the primary resource for rubric construction. During the first phase of this study, the goal was to obtain student performance on the task in which students were to display their IC. The performance was analyzed through the lens of conversation analysis (CA) and applied conversation analysis to provide a structural explanation on how student participants organized their interactions in order to complete the role-play, and how higher competent students handle this social activity compared to lower competent students in this dataset. The summary of the qualitative findings in Phase I was organized into a rubric, which was implemented and analyzed in the quantitative phase of the study.

Phase I: Qualitative analysis

This qualitative phase is designed to address RQs 1-3: “From the students’ roleplay performance data, what are the interactional phenomena in the form of actions or courses of action which can be established as the targets for comparing IC across the dataset?”, “What are students’ methods, in varying degrees of success, in accomplishing the actions or courses of

action identified as the targeted interactional phenomena?", and "How can the rich description of students' task performance inform the data-driven construction of the IC assessment rubric?"

Students' performance on the Socializing task was video recorded for the current research and analyzed using qualitative methods of conversation analysis (CA). The data collection in this phase of this study sampled student test taker performance in February 2016 during the first semester of the same academic year.

Student participants. One hundred and eighty students participated in this study and 34 group oral roleplay performances were collected through the help of the course coordinator of the course who has also been part of the material and curriculum development from the start. These students were second-year and third-year undergraduate students from different engineering majors offered at this university: Computer Engineering, Mechanical Engineering; Industrial Engineering, Civil Engineering, Chemical Engineering, and Nuclear Engineering. One hundred and sixty student participants were male (approximately 89%), and 20 were female. This ratio represents the current demographic of the engineering job market in the country. Descriptive statistics of their scores based on the original rubric are shown in Table 4.1 below. Across all four criteria, students were rated quite high. For three of the criteria (group collaboration, content, and delivery), means were more than four out of five, with a small spread around 0.5 across all four criteria. The most difficult criterion from the original rubric was the language and pronunciation criterion, followed by delivery, content, and group collaboration. In terms of distribution of scores, group collaboration and content scores were negatively skewed, meaning that more students scored above the means on these categories. The language and pronunciation category appeared to be the most normal. The delivery scores, on the other hand, appeared slightly leptokurtic, which means that many students scored at the mean, resulting in a distribution that is taller than normal.

Table 4.1
Descriptive Statistics of Participants' Scores on the Original Rubric

	N	*Min*	*Max*	*Mean*	*SD*	*Skewness*	*S.E.S*	*Kurtosis*	*S.E.K*
Group collaboration	180	3	5	4.20	0.53	-0.493	0.18	0.223	0.36
Content	180	3	5	4.17	0.55	-0.434	0.18	-0.043	0.36
Language and pronunciation	180	2.5	5	3.88	0.56	-0.169	0.18	-0.245	0.36
Delivery	180	2.5	5	4.02	0.60	-0.199	0.18	-0.557	0.36

The performances of 34 group interactions of the students who gave their consent to participate in the study were video recorded. The group sizes ranged from four to six students per group. Video data were deemed essential in this study since embodied actions are an integrated part of IC in co-present interaction that would not be available in audio recorded data.

Data analysis. The video data were reviewed and transcribed following conversation analysis (CA) conventions (Atkinson & Heritage, 1984) to reveal the detail how students organize their turns and the sequence through which they organized their actions during their roleplay performance.

The reasons why CA was chosen to provide an analytical description of how students perform the roleplay are as follow. First, CA is uniquely conceptualized to capture the organizations and structures of social interaction in both of its procedures and outcomes. Second, it takes into account the co-constructed nature of talk-in-interaction, the manner of which can be used to provide a specific description of how co-construction can be operationalized for language testing and assessment of oral communication skills. With CA, the questions of what has been individually achieved or collectively achieved, by what resources, and in what manners can be explicitly addressed. Thirdly, because CA takes an emic perspective in analyzing what and how actions and accomplishments have been made in interaction, it can be illuminating to ascertain how student accomplishments match with those expected by teachers and users of test outcome. Finally, adhering to the principles of using CA as a lens for qualitative observation will help us steer away from treating IC and co-construction as fixed patterns of interaction which reduced IC

into a certain kind of products which rid IC of its rich procedural component that we also need to measure.

The goals of the qualitative analysis are to identify recurring actions or courses of action throughout the roleplay data, which can provide a consistent observational basis for the assessment of IC. Given the fine granularity which actions and courses of action can differ, to establish a common target for comparison of student IC, this study is attempting to adhere to the principles for quantifying talk-in-interaction from Schegloff (1993), which have been adapted into practical recommendations for longitudinal developmental studies summarized by Pekarek Doehler et al. (2018) and Pekarek Doehler and Berger (2018).

This fine-grained qualitative observation of interactional accomplishments and interactional resources making such accomplishment possible will lay a foundation for constructing a measurement scale for assessing IC in the subsequent stage. As Bond and Fox (2015) note on the nexus between qualitative observation and quantitative observation, all observation starts off qualitatively, and counting starts off with repeated observations before those observations can be quantified. Actions or courses of action identified as the target interactional phenomena for the assessment task will be treated as items against which each person's performance will be coded for success or failure for each item under the rubric guidelines.

Phase II: Quantitative analysis

With the findings from the first phase of this study, we now have a set of descriptors which were organized in the form of rubrics or rating scales for the second phase of the study. The data collection and analysis in Phase II was designed to answer RQs 4-6: "Given the proposed rubric for assessing IC in this roleplay task, how reliable are the raters in applying the

scale to rate the students' performance?", "Are there any detectable biases in how the raters apply the scale to rate the students' performance?", and "To what extent can the current study argue for the validity of the proposed rubric and rating scale for assessing IC in this context?".

Raters. Six experienced language teachers volunteered to participate as raters in this study. Two of the teacher raters were current instructors of the course *Communication and Presentation Skills.* Two of the teacher raters were recruited through the University of Hawai'i at Manoa's Conversation Analysis group mailing list. The last two teacher raters were recruited through word of mouth via the connection of language teachers teaching ESL in Hawai'i. Among this group of raters, 50 percent of them were male, and the others were female. Two of the raters were native speakers of English, and the other four were native speakers of Thai, Chinese, German, and Korean. A summary of the raters' background experiences and their assigned identifications for current study is presented in Table 4.2 below.

Table 4.2
Raters' Background Profiles

	Gender	English L1	English L2	CA training background	ESL teacher in Hawai'i	Original Instructors
Rater 1	f		✔	✔		
Rater 2	f		✔			✔
Rater 3	m	✔				✔
Rater 4	m		✔	✔		
Rater 5	m	✔			✔	
Rater 6	f		✔		✔	

All six raters underwent an individual training with the researcher which lasted two hours to familiarize themselves with the rubric and practice identifying the targeted actions and rating excerpts of different activities with the researcher before being asked to rate one whole performance to check that they were ready to carry out the ratings of the remaining roleplays on their own.

Data screening. Among the 34 group performances of 180 students, there were 32 individual cases in which the raters could not produce a rating in at least one of the action items

due to poor video and sound quality, or that the video angles made it impossible for the raters to discern speakers' contributions to the group role-play at one point throughout the student interaction. This is quite a challenge for a research study using peer group role-play format given that participants were free to move around and form smaller subgroups during their talk at different points of their interaction, and a single set of camera and voice recorder could not sufficiently capture clearly all interactions that occurred during the group talk.

For this reason, the missing data were eliminated from the analysis case-wise to ensure that the remaining pool of student performances were eligible cases in which all raters agreed that they were assessible for the entirety of the role-play. This reduced the current number of participants from 180 students to 148 students with completed ratings on all the action items across the six raters.

Data analysis. For this study, we opt for using item response theory to guide our analysis of student accomplishments. Multi-faceted Rasch partial credit response model will be used to analyze the probabilistic relationship between student interactional ability, the characteristics of each criterion (items), and the characteristics of each rater in a three-facet model. The multi-faceted Rasch model can consider the contribution of each item and each rater to explain the overall variance of the students' scores, then weight each of their contributions in making a prediction about the person's true ability. We can check the fit statistics to determine how much we can trust the result generated from this model.

Multi-faceted Rasch analysis can also reveal any potential interactions between different facets, which can point to possible bias patterns between raters and items, or raters and different examinees. All in all, through the lens of the rubric we have constructed, we can explore both the construct of IC and also the rater behaviors when applying the rubric to assess IC.

Before analyzing the scores generated from the proposed new rubric, the data was checked to make sure that it met the assumption of unidimentionality required for multi-faceted Rasch partial credit model. It is very important for any models under item response theory (IRT) that all items within a test must be assessing one single construct and therefore has only one dimension. The combined scores from all the raters for each category were checked if they were unidimentional through principle component analysis (PCA) using the SPSS program. If PCA revealed that the action items did not fall under a single construct, analysis using multi-faceted measurement model would be performed separately for each construct.

To provide evidence of score reliability and validity of the rubric and the construct of interactional competence (IC), the primary purpose of conducting the FACETS analysis is to study the characteristics of the students, the raters, the target actions (items), and the scale used in rating the student performances.

Because of the use of a rating scale, the Partial Credit Model (Masters, 1982), an extension of the Rasch measurement model which deals with graded scoring often used in performance-based tests, was employed. Through the many-facets Rasch model (Linacre, 1989), we can study students' performance while simultaneously taking into account the properties of the tests and the raters all at the same time (Bond & Fox, 2015). This single framework allows us to capture complexities within a measurement situation, as not only does it provide an estimate of students' ability, adjusted for task difficulty, and raters' severity, but it also provides an analysis of interactions between these different facets which can identify issues from raters' behaviors or the rating scale for further analysis.

Finally, based on the findings synthesized from both qualitative and quantitative phases, the study will attempt to make an argument for the validity of the new rubric generated in the study in capture IC in performing socializing roleplay task.

CHAPTER 5

A MICROANALYSIS OF STUDENT ROLEPLAY PERFORMANCES

A part of the original rubric for this *socializing task* is the "content" of the interaction. Under this category, which was worth 25 percent of the original total score aside from the other categories which included language, delivery and group collaboration, the rubric reads as follows: "Content (5 points) - Talking about personal information, jobs and responsibilities; Talking about your company (accurate information), Business card." This original rubric was made available to the students early on as it was published as part of their textbook materials, produced in house by the instructors especially for this course.

In this qualitative analysis phase, the study explored evidence of student interactional competence (IC) displayed on the selected interactional activities that were compulsory in carrying out the roleplay task. To maximize comparability of students' interactional performances, the study narrowed down the analytical focus to eight social activities following the overall sequential organization constitutive of how students managed the *socializing task*'s completion in the dataset. The framework of applied conversation analysis was used in guiding the analytical findings and interpretations in this chapter. This chapter provides structural explanations on how higher competent students handle this social activity when compared to lower competent students on each of the identified interactional activities representing this task.

A rubric that is generated as a product of this qualitative analysis will inevitably embody the construct of IC. Selecting the interactional activities also entails prioritizing certain interactional styles, methods, and interactional outcomes that the test deems more desirable, effective, and "normal". Unavoidably, the test overlooks some other aspects of interactional

achievements or competencies in the process. The consequence of such selection is another empirical question that should also be addressed, but it should be done in future projects.

Through inspecting the overall sequential organizations of the roleplay performances, there were five recurring activities which the majority of the participants consistently performed and therefore warrant our selection of these actions to be included in the grading rubric. The activities selected for the rubric are divided into two main sections: action production and recipient actions. On the production side, the study included five activities: (a) self-introduction, (b) talking about their company and job responsibilities, (c) doing contact exchange, (d) making post-conference plans, and (e) negotiating task termination. For recipient actions, three aspects of recipient designs were included based on the frequency of their occurrences: (f) understanding display, (g) alignment management, and (h) maintaining affiliation. The content in this chapter is organized into six sections for each of the production activities and one section discussing recipient actions.

Self-introduction (SI)

In a multiparty talk on this roleplay task, students' managements of their self-introductions (SI) typically comprises of three steps: invoking SI through topic transition, completing their SI and nominating the next speaker or closing the SI activity. Interactionally competent candidates should be able to display that they can construct their SI competently with their co-participants, with appropriate turn designs that fit each action to its audience and thus show that they know the sequence organization of this activity. Beyond their knowledge and ability to execute their SI, they also have to be able to mend their course of actions through effective employment of repair if and when any potential understanding problems occur. As these problems are a natural part of interaction and can hardly be planned or predicted in

advance, students' ability to manage repair practices is ingrained throughout all actions and activities under this test task.

Let us begin by comparing two student SI performances in the first two excerpts in order to see the contrast in observable aspects of IC between stronger and weaker candidates. Before invoking SI in their respective talks, students in both examples do the roleplays with an exchange concerning the time and location of the conference. In excerpt 5.1, Sutham, whose role was a software engineer from a small e-learning company, managed a transition from ongoing activity into his SI in a gradual manner.

Excerpt 5.1 *Sutham's SI*

```
76      Sut:   a:::nd (.) here we are!
77             the::: welcoming party.
78             (.7)
79→     Sut:   but- (.) ↑by the way we haven't introduced ourselves yet,
80             maybe (.5) I going to introduce myself +first=
        Pan:                                          +nod

81→            =my name +is Sutham (.5) or >you can call me North<.

82             +(.4)
        oth:   +nod

83      Sut:   <I'm from> (.2) New Zealand.=
84             =But now I working at (.4) Khan Academy=at U-S.

85             +(.5)
        Sut:   +nod

86      Pan:   +O:h.
        Pan:   +small upward nod

87      Tho:   +North yes? [North.
        Tho:   +lean over, GZ→Sut

88      Sut:               [North
```

After a long pause in line 78, the previous activity had come to a potential closing, leaving an open space for anyone to take the next turn. In line 79, Sutham secured a turn with "but" with a cut off and a micropause. Then he did a restart with a transition marker "by the way" making explicit that a change of topic or a new course of action is forthcoming. After the transition work, instead of going directly into his SI, he first made a comment on how the group's SI sequence was missing, then with a pause and hesitant marker "maybe" in line 80, he

tentatively volunteered to be the first one and managed to secure a go-ahead signal from Panu which come overlapping the end of his turn in line 80.

Sutham's work to project his upcoming SI, though still arguably not target-like, shows a passable attempt to transition from prior activity, getting to the conference venue, to create an opportunity for his SI. He also showed his orientation to his SI initiation as a joint activity as he was able to secure ratifications from Panu (line 80) and other group members (line 82) along the way.

When we compare Sutham's SI initiation with Mark's SI initiation in the second excerpt, a stark contrast can be observed. In this group's roleplay, students also made use of an inquiry into the starting time of the conference in the task opening sequence. We can see a question-answer sequence between line 1-2, with a minimal post-expansion sequence in line 3-5 completing this task opening. To inspect this group's transition into their SIs, we can start with Mark's turn design in line 6 below.

Excerpt 5.2 *"I'm… I'm Mark"*

```
01     Ben:   uh hi sorry do you know when does the expo start?

02     Mar:   uh +(.) +the next hour.
       Mar:      +GZ shot up
                       +GZ→Ben

03     Ben:   ↑oh really.=
04     Mar:   =˚˚yeah˚˚

05     Ben:   o[kay_
06→    Mar:    [hel- hello. I'm-=
07     Ben:   =nice to meet you

08     Jam:   +>hello<=
              +offer handshake to Max

09     Max:   =[>hello     ]
10     Ben:    [nice to me-] [nice to meet you.

11     Jam:                  +[˚(hey)˚
       Jam:                   +offer handshake to Tan

12     Mar:   +I'm (.) I'm mark,
       Mar:   +RH point to himself
```

First of all, a transition from the question-answer sequence (lines 1-5) leading up to the SI action sequence was noticeably missing. In line 6, Mark initiated a greeting in an overlap with

Ben's "okay" in line 5 and naturally had to restart his "hello" after the overlap. It could be argued that Mark's "I'm" at the end of his turn in line 6 resembles an onset of a self-introduction; however, it is unfortunate that we could not see how he would completed his SI here as his turn was cut short by Ben's utterance "nice to meet you" in line 7. This uncertainty which led to apparent conflicting interactional goals among the participants was arguably contributed by the absent of transitional work that would otherwise be expected by either Ben or Mark between line 5 and 6. The consequence was an abrupt break between the previous question-answer sequence and a new one, leaving visible interactional problems which participants also failed to address by any repair sequence. The participants at that point abandoned the SI activity and continued in a greeting sequence (lines 8-11), except Mark, who did not engage in the greeting sequence, but simply reinitiated his SI again in line 12, issuing it as a topic announcement (Button & Casey, 1984).

Mark's SI performance showed problems in sequence boundary awareness which led to two missed opportunities to project his upcoming SI via topic transition. In short, the organization of actions for this group reveals many interactional problems which raters may use as evidence for students' lack of awareness of sequential organization as well as poor turn designs in initiating their SI activity.

Another challenge of completing SI competently in a multiparty roleplay such as this one also resides in how students complete and manage the transition away from their SI with other group members. With this task, we can distinguish students with varying degree of IC by observing how they manage their SI closings. In excerpt 3, our focal participants are Panu and Tara. Sequentially, once one's SI is completed, the next relevant action would be an exchange of "nice to meet you" or any other positive assessment tokens and a nomination of the next speaker to continue or to initiate a new conversation topic. When we compare Panu's and Tara's methods

in nominating the next speaker, we can see a difference between the two methods that the two mobilized.

Excerpt 5.3 *Panu and Tara's SIs*

```
105    Pan:   my name is Panu (.) (last name),
106           I'm from England. (.7) um (.8) I'm::: currently
107           I'm working (.4) for an estate company.
108           (.3)

109→   Pan:   how about +you:.
       Pan:              +RH→Tar

110    Tar:   +↓oh
       Tar:   +visibly drew in a breath, GZ up

111    Pan:   what is your name and:: (.) where're you from.
112    Tar:   um my name is Tara (.2) I from Russia.
113           I am (1.1) I-T supporter.
114           (.3)
115    Tar:   ˚yeah˚.
116           (.4)

117    Pan:   [˚ah˚ +nice to meet you.]
       Pan:          +bow

118    Tho:   [I-T supporter          ] oh [you
119    Tar:                                [yeah.
120    Tho:   you have come a long:: way.
121    Tar:   ↑ye:ah.

122→          +(.5)
       Tar:   +turn to Tan, RH→Tan

123    Tan:   +Oh
       Tan:   +RH point to himself, GZ→Tar/Tho

124    Tar:   um

125           +(.4)
       Tan:   +cont. RH point to himself, GZ→Tar/Tho
       Tar:   +GZ→Tan, nod twice

126    Pan:   ˚how 'bout you˚.
```

In the case of Panu, he began his SI in line 105 and appeared to have completed it in line 107 after having mentioned his name, home country and a rough description about his job. In line 109, after a brief pause, Panu moved to assign Tara as the next speaker. He initially formulated a question "how about you" (line 109), which was accompanied by his hand gesture, clearly selecting Tara as the next speaker. He then preemptively self-repaired his question (line 111) after Tara's upcoming response was slightly delayed, making his action, from the one with higher degree of implicit reference to two separate questions which were more explicit.

In the case of Tara, we first observed that he proceeded to answer both Panu's questions (his name and where he was from) in line 112. Then, he displayed his recognition and alignment of the activity to be beyond simply answering the questions by adding that he worked as an IT supporter (line 113)–mirroring the pattern of SI covered by other group members so far. It should be noted that with Tara, the next relevant action after his SI was noticeably missing. After a brief pause, he softly uttered 'yeah' with a falling intonation, relinquishing his turn. Then, there was a brief pause which Panu and Thor both oriented to as an indication of Tara's being done with his turn. Both of them competed to launch their post expansion sequences: minimal in Panu's case (line 117) and a little more extensive in Thor's case (lines 118-121). Reaching another sequence closure relevant point, Tara did the embodied next-speaker selection (line 122) by turning to face another participant, Tan, and point his right hand in Tan's direction. Tara's method of nominating the next speaker was visibly problematic as we could see from Tan's reception of his action. Tan was visibly unsure about Tara's nomination of the next speaker. He shifted his gaze between Tara and Thor as he pointed his right hand to himself (line 123), interpretable as a nonverbal request for verification that the turn was actually his. Although Tara noted and recognized the trouble, he again resorted to a non-verbal solution to supply Tan the repair (line 125), failing to demonstrate the variety and range of his resources in interaction, particularly in closing his SI.

From the analyses of the above excerpts, students who were demonstrably more competent in their management of SI in this task were those who showed awareness of both sequence organization and the overall sequential organization of a typical SI and had better control over their execution of their actions. In the sequential environment at the onset of their SIs, competent students are those who are able to manage the transition smoothly away from the prior activity and create an opportunity for their SI. Once they have completed their SI, students

would have to demonstrate that they are able to actively take part in progressing the roleplay forward, either by nominating the next speaker, initiating an expansion of the topic at hand or initiate a sequence closure before transitioning to the next activity on the task.

Another viable consideration to distinguish students at various command of IC in doing SI is in the quality in which students composed their actions. A consideration concerning the extent to which students' understanding of the target situation is reveal through how they construct their roles in interaction. To this end, membership categorization analysis (MCA) can provide a framework through which raters can judge a stronger composition from weaker ones.

Performing the roleplay, students were intrinsically asked to display their understanding of the target situation. How closely their version resembles the "real" situation is on display for raters to assess. It is unfortunate that the roles this task is asking the students to play are considerably far removed from their usual environment. Being college students without much if any working experience, it could be difficult for them to play a role of working adults representing a company at a trade/business conference socializing and forging professional networks. However, there was still evidence of recognizable membership knowledge that this group of students displayed, at varying degrees, to which raters can pay attention.

Membership knowledge is manifested in the way members do things or perform actions in a way that are recognizable as category-bound actions. In excerpt 6, one apparent observation in the way Takeshi displayed his membership knowledge in his role as a conference organizer was in line 71 where he inquired of other participants what they thought about the conference, an action which is not integral to a typical SI sequence had his role been something else. Because of his chosen role, his audience, as well as the raters, were permitted a reinterpretation of his SI sequence, not as a core activity in itself, but an opening sequence before launching the main action that is to check if conference-goers were satisfied with the organization of that event. As a

result, the lack of any topic transition prior to Takeshi's initiation of his SI was retrospectively accounted for by his announcement that he was part of the organizing staff of the conference. Although the design and composition of Takeshi's SI was not problem-free, we can recognize that the placement of his SI sequence before the 'main business' action in line 71 was congruent with his role as an event organizer, thus provided an evidence, to some degree, of his membership knowledge and IC.

Excerpt 5.6 *"This event?"*

```
52            (.3)
53     Tak:   hello everyone
54            (.4)
55     Tak:   my name (is) Takeshi
56     Bor:   Take[shi
57     Bob:       [Takeshi=
58     Tak:   =yeah I'm I'm organizer and organize (.4) this event,

59     Bob:   +this [ <$event$?>] [H h h h
       Bob:   +RH point down

60     Bor:          [this event?]
61     Tak:                       [oh↑ yeah.
62     Ben:   [oh↑
63     Bob:   [Okay?=
64     Tak:   =my (.5) my head office uh in (.7) Beijing but

65            I come from (.6) uh Japan,+(.4)>Hokkaidoyouknow<?
       Bob:                               +nod

66     Bob:   [Hokkai+do] yes I +know that.
       Bob:           +nod & snap his fingers

67     Bor:   [Hokkaido ]

68     Ben:   [+yes     ]
       Ben:    +nod

69     Bob:   I like it.
70            (.4)
71     Tak:   um how (.) how about this expo?
```

Components of IC in any given speech community are defined largely by the procedural knowledge required by a competent member to manage their mundane businesses. Such knowledge can be revealed through the compositions of their actions and each of the turns making up into sequences of the target actions. For this SI activity, the above descriptions of competent and problematic performances are to highlight what students were able to accomplish interactionally in what seems as a simple an activity as introducing oneself. Students who can

manage the activity competently deserve to be recognized for their skills, while students who struggle to manage such activity can learn from their mistakes in a way that could improve their future interactions.

Work Talk (WT)

The next topic that students often moved onto after completing their SI was to share with their group members what their company does, what their job titles and responsibilities are and also what their specific goals or purposes they wanted to achieve at the conference. Altogether, this set of topics will be discussed as "work talk" (WT). Given the design of the task, this shift of topic into work related one theoretically would require IC to manage not just the topics, but also their stances in relations to other co-participants and intersubjectivity among group members as the roleplay unfolds. The students were instructed to do research about their chosen roles, preparing relevant information in order to interact with their peers in the roleplay task. Students with access to such knowledge were able to tap into the prepared resource as they managed both their topic talk and stance work while participating in the roleplay.

In determining the exemplary cases for assessing WT, we compare student performances in two sequential environments of their WTs: topic initiation and further expansion of their WT. The analysis below presents descriptions of competent and problematic cases of each sequential environment of WT in order to serve as baseline models which could guide raters in reaching their scoring decisions.

At the onset of WT, it is rare to find a transition managed seamlessly from any prior topic into what was recognizable as WT in this dataset. Students overall resorted to initiate the WT activity following a substantive silence which occur succeeding a closure of the previous topic.

However, there are some discernable qualities between different levels of IC in the designs of their actions to be discussed below.

First, let us compare two excerpts (Excerpts 5.7 and 5.8) of how stronger and weaker candidates manage the initiations of their WTs.

Excerpt 5.7 *"I'm totally different"*

```
21      Hug:   that's good country↑
22             (1.4)
23→     Hug:   and:::: (.2) what company do you work for?
24      Jon:   um. I work for chevron.
25      Hug:   Chevron?
26      Jon:   +how about you.
        Jon:   +RH open in Kai's direction, GZ→Kai

27      Kai:   I work in Mitsu oil company.
28      Hug:   +Mitsu oil=and:::
        Hug:   +nod

29→     Hug:   ↑+I'm totally different because I work in I-T
30             and electronic (.3) ah::: company,=
31             =called Xiao Mi.
32      Kai:   Xiao Mi?
33             (.3)
34      Hug:   Xiao. Mi.=
35      Kai:   =Xiao. Mi.
36      Hug:   yeah=
37      Kai:   =+yeah.
                +nod, GZ→Hug
```

Excerpt 5.7 exemplifies a stronger candidate. Following a topic termination, Hugo displayed an orientation to maintain the flow of conversation after their SI activity had come to a close. In line 23, he started the turn with "and," which is typically used to mark a syntactic continuation between the upcoming turn and the previous turn, reopening the potential topic closure which was projected previously. Despite the initial attempt to connect WT with the prior talk, Hugo's question asking about one's company (line 23) was still hearable as a new topic as the object of his inquiry had shifted to a work-related matter which had little topical coherence with the previous talk about one's home country. Hugo also oriented to this topic restart in that he contrasted an elongated "and" before the micropause (line 23) with the stream of question delivered after the micropause. With this initiation, Hugo selected Jon as the first person to develop his WT, in which responded and quickly passed on to Kai soon after (line 26). After an

acknowledgment token in line 28, Hugo shifted the conversation focus back to himself. The design of his turn in lines 29–31 offered a chance for us to see his competency to do more relational work through his WT. Instead of merely stating information to make up his WT, Hugo first offered an upshot of the upcoming content at the beginning of his turn ("I'm totally different") which showed his acknowledgment of other students' job positions and also highlighted the relationship between the industries he and others were from.

Next, we move on to consider the management of WT initiation in weaker students. Excerpt 5.8 also began at the closing of SI sequence when Thomas and Josh exchanged their "nice to meet you" tokens (line 27) and a handshake over a very extensive pause in line 28.

Excerpt 5.8 *Kate's WT initiation*

```
27      Tho:   nice to meet you. too.

28             +(4.8)
               +Tho and Jos shake hands

29      Tho:   um::
30             (1.9)

31→     Kat:   +what do you +do.
        Kat:   +GZ→Yos       +GZ→Tho

32             +(1.8)
        Kat:   +GZ→Tho
        Tho:   +GZ→Kat, nod twice


33      Kat:   +what do you do?
        Kat:   +GZ→Jos, RH point to Jos

34             (.)
35      Kat:   [Jon?]
26      Jos:   [uh  ] (.2) I'm a (.7) my (.) company
27             uh (.4) making uh (.7) aircraft manufacturing
```

Following that pause, we can see that both Thomas and Kate were equally hesitant to take the turn. Thomas actually verbalized a stretched-out filler ("um") followed by another stand-still long pause that he seemed to share with Kate who later broke the silence with a question inquiring about someone's job in line 31. The formulation of Kate's question was hearable as awkward given that it did not follow any transition device. Her accompanying eye gazes in line 31 also made it unclear whom her intended audience was as she first directed her gaze at Yoshi

at the beginning of the turn and then shifted her gaze to Thomas towards the end of her turn. This resulted in another lengthy silence during which we can see nonverbal insertion sequence between Kate and Thomas (line 32)–Thomas looked back at Kate, whose gaze remained fixed at Thomas, who then issued small head nods in a move which is readable as a go-ahead signal. We then see that Kate reissued her question in line 33. This time, her embodied gestures helped aid the formation of her action, selecting one clear audience for her question as she directed her gaze solely at Josh and used her right hand to also point to him. It is also notable that she did not change the form of her question at all, except the added rising intonation at the end of the question. A micropause which followed her turn in line 34 prompted Kate to repair her action one more time. Still perceiving that audience selection was the trouble source, she used another method for recipient selection, this time using her recipient's name with a rising intonation (line 35). This may have been a misjudgment in identifying the trouble source on her part as we see that her repair in line 35 overlapped with the onset of Josh's turn, knowing it was his turn, in line 36 starting his WT in a delayed formulation.

The insertion sequence taking place during the silence portion of the roleplay often reveals how students with limited linguistic resources mobilize their IC in coping with the roleplay task requirements. In this case, Thomas and Kate during the insertion sequence (line 32) were trying to work out which of them should take the turn to initiate the group's WT. Thomas yielded and Kate took the next turn to issue the initiation. For the purpose of this assessment task, the coping interactional competency through the use of embodied gestures which facilitated Thomas and Kate's turn-taking practice was not directly relevant to the construct. However, it could be worthwhile for future studies to explore this kind of task-related interactional competency which students can tap into when they have to cope with the gap between the task requirement and their L2 proficiency.

As these two excerpts show, one key difference in how participants at different levels of IC managed the initiation of their WT lies in the design of their sequences. Stronger candidates always display a recognition of topic or activity boundaries and would attempt to provide transition markers before they commence their WT. Stronger candidates can also fine tune the design of their action in relation what was known about other participants up to the point that they took their turns. Weaker test takers, on the other hand, struggle with managing the transition. They also tend to be inflexible with their turn design, reflecting their impoverished set of linguistic resources, and may resort to shift their participation frameworks from participating in their assumed roles to participating as a student trying to complete the task.

While both SI and WT require students to manage topic initiation or transitions at the onset of the activities, for WT, the task requirements go further to demand more extensive development of the WT topics to cover all of the required areas of WT: talking about their companies, their job positions and the responsibilities they had in their respective companies. To comply with these requirements, students must co-navigate the WT with their peers, and having to do so could impose a greater challenge to students as they had to manage their WT in ways that are sequentially appropriate while also display their understanding of how the interaction in their respective roles would unfold in real-life situation outside of the assessment context.

After successfully launching WT, the sequence environment which hinges on whether topic development would take place is the post expansion sequences. Almost like a game of hot potato, students in this dataset often quickly passed on their turns to other group members after their first question-answer sequence and thus abandoned a chance to fully developing their WTs.

To illustrate, in Excerpt 5.9 below, we can see a competition between turn passing (Hugo) and topic development (Kai) in a talk with Ann as the focal participant.

Excerpt 5.9 *"What do you… and what do you do?"*

```
160    Chr:   what about you (lady)?
161    Ann:   oh I work in NASA.
```

```
162    Chr:   ↑<N A S A>.
163    Kai:   NASA?
164    Hug:   NASA
165    Jon:   [it's cool.
166    Kai:   [˚Ooh˚.
167           (.5)
168    Hug:   it's a big organization.
169    Ann:   yes.
170           (1.6)

171→   Kai:   [+what do you-
       Kai:    +GZ→Ann

172→   Hug:   [+and what about you?
       Hug:    +GZ, RH → Chr

173           (.2)
177    Chr:   um
178    Hug:   ↑oh you say=
179    Chr:   =£from Apple£ ↑yeah.
180    Kai:   [↑oh::
181    Hug:   [I'm sorry for that
182           (.5)

183    Kai:   and +(.) what do you do?
       Kai:       +RH to Ann

184           ˚in NASA˚.
```

Before we further discuss the excerpt, it should be noted that Ann's participation in this WT activity was rather passive. Her ability to sustain her WT in this roleplay was only possible thanks to her peer's initiation and her on-target activity alignment, which nevertheless showed that she was able to recognize the overall sequential organization of the roleplay and manage her participation accordingly. Now, regarding the development of her WT, after the initial sequence of her WT (lines 160–169), there was a long pause in line 170 which made the prospective trajectory of the activity rather uncertain. Breaking that silence, Ann's co-participants, Kai and Hugo, both launched their projected next action at the same time, resulting in an overlap between lines 171 and 172. Kai wanted to issue a post-expansion turn to further develop Ann's WT. Hugo oriented to the pause in line 170 as topic termination of Ann's WT and looked to Christian to launch his WT next. Hugo's projection of the next activity in the roleplay had won out as Kai did not finish his sentence and also it was Christian who provided a response to Hugo's initiation in the turns that followed. However, when it turned out that Christian was not a viable choice as the

next speaker (Christian had already conducted his WT prior to this excerpt), Kai found another chance to pursue the topic development for Ann's WT as he re-issued and this time accomplished, his post-expansion sequence initiation again in lines 183 - 184.

One thing raters should note is that even though Ann was the focal participant in this excerpt, other group members were also active in their participations to further develop Ann's WT, taking on a responsibility to initiate expansion sequences to sustain topic talk with other group members. This active participation should also be awarded, regardless of whose WT was in focus, as long as their topic expansion initiations were done in a non-disruptive and sequentially appropriate timing and the content were cohesively designed to fit the ongoing WT talk. To illustrate how the non-focal participants could initiate topic expansions successfully in interaction, Excerpt 5.10 shows a continuation of Ann's WT.

Excerpt 5.10 *"Sound problem?"*

```
183→  Kai:   and +(.) what do you do?
      Kai:       +RH to Ann

184           ˚in NASA˚.
185   Ann:    um >I am< research aerospace engineer.
186           (.3)
184   all:    um[:::.
185   Chr:      [aerospace engineer.
186   Hug:    so (.) you are po- your position is a:: engineering?
187   Ann:    yes=
188   Hug:    =yes=ah [engineer
189   Ann:            [I'm a engineering.=
190→  Kai:    =what's your responsibilities?
191           (.7)
192   Ann:    um (.5) my job is (.) about create and develop spaceship,
193   Chr:    ↑um[:: ]
194→  Ann:       [and] (.) also include about working about
195           sound ↑program.
196   Hug:    sound program?
197   Chr:    ˚sound problem?˚
198   Ann:    we make (.2) astronaut safe from ↑danger.
199           when they ↑working.
200   Kai:    oh!=
201   Ann:    =that's [my job.
202   Hug:            [↑um.
```

In Kai's first successful attempt to further the group's WT with Ann, his formulation of the question "what do you do in NASA" (lines 183-4) elicited her response about her job title. There was a confirmation request insertion sequence initiated by Hugo in line 186 which

generalized her job title from “research aerospace engineer” to simply an engineer, to which Ann confirmed the category (line 187) and re-verbalized this less specified job title (line 189) to close the sequence.

In the next iteration of expansion sequence, we can also see in line 190 that Kai formulated another question explicitly soliciting Ann for her job responsibilities in line 190. Knowing that this is one of the main target matters test takers must cover to fulfil the task, perhaps Kai had issued this question as a self-repair given that his first and less specific, formulation was not able to prompt Ann to talk about her job responsibilities.

Recognizing and capturing a chance to bring up one’s WT during an ongoing talk is one challenge. What students with lower level of IC often struggled is also how to organize the content they prepared naturally in the roleplay once they had their turns to talk. What we want to see in competent construction of WT is that it was organized naturally as part of an ongoing conversation. This means that students must issue their WT content in talk, not as a monologue.

A salient indicator that students were orienting to any part of the roleplay as a monologue is in the extent to which students were in control of intersubjectivity. In other words, do they recognize any potential understanding problems while they were having a turn? Or, to what extent do they pay attention to the group members’ displays of understanding of what was going on in the roleplay? Ann’s performance shown in excerpt 12 also offers an example of an interaction in which participation attempts from others were ignored. More discussion of Ann’s failure to maintain intersubjectivity in this excerpt can be found below when we discuss understanding display as part of the construct of recipient actions.

In another episode of WT taken from the same group, Excerpt 10 provides two cases of successful management of WT by Christian (lines 245–273) and, later on, Hugo (lines 274–292). Interactional skills that these two were able to reveal through their construction of their WTs

include (1) an ability to recognize the sequential environment for WT development and take an active role deploy their WT accordingly and (2) a capability to situate their WT in the ongoing interaction while also maintain intersubjectivity with other group members.

Excerpt 5.11 *Christian and Hugo's WT*

```
245    Hug:   +and how about you?=
       Hug:   +RH to Chr

246    Jon:   =>+how about you?<
       Jon     +GZ→Chr

247    Chr:   ah I'm a software engineer.
248    Hug:   u::m=
249    Chr:   =at Apple.
250           (.5)
251    Chr:   a::h (.2) my main duty (.) is:::: to::::
252           develop some new technologies.
253           for:: ↑um↓ for people.
254           to live:: more com↑fortably and easy.
255           (.4)
256    Chr:   ↑ah to maintain:: the ai-phone operation system.
257           known as ai-o-es.
258           (.2)
259    Chr:   an::d↑ to::: (.) invent a new ↑product.
260           once in a while.=
261           =to keep the company in spotlight.
262           (.4)
263    Hug:   [um[: : :.
264    Jon:   [um[: : :.
265    Chr:      [>what about you Alex?<
266    Ale:   so is so (.3) if you- I want to- (.) get some application
267           can you do it for me?=
268    Chr:   =↑yes! you can a:sk me.
269    Kai:   [wo::::w
270    Hug:   [wo::::w
271    Chr:   let's talk about this later.
272    Ale:   [o[kay.
273    Chr:     [after the conference.
274→   Hug:   despite of (.) we are working in the: a:h smartphone company,
275           but my duty is totally different from you,
276    Chr:   [u::m   ]
276    Hug:   [because] my duty is to: like ah:: (.)
277           to:: (.2) operate the marketing,
278           (.3)
279    Hug:   ah outside of China↑=like India,
280           Singapore or:: Malaysia,
281           (.5)
282    Hug:   ˚that's˚
283           (.3)
284    Chr:   you want to make your company world-wide,
285           (.2)
286    Hug:   ah:↑ yes. (.2) in the process.
287           (.3)
288    Hug:   it's in process.
289    Chr:   ↑um::.
290           (.5)
291    Ale:   okay
292    Hug:   and how about [you?
293    Ale:                 [for me.
```

```
294          (.3)
295    Ale:  I'm electrical engineering.
```

Another quality of Christian's and particularly Hugo's formulations of their WTs is in the professional narrative that they were able to construct surrounding their companies and their job positions. Since the students all prepared their roles individually and therefore would have no knowledge about whom they eventually would be partnered up with, it is an added challenge if they were to be in the same group with participants from the same industry (market competitors) or the same company (job position relationship or hierarchy). For Hugo and Christian, they were both from mobile phone technology companies. Christian was a software engineer from Apple, a large and well-known corporation based in the US, and Hugo was a vice president from Xiao Mi, a China-based company which is significantly newer and therefore much lesser known than Apple. In the excerpt, we can see that both of them were able to construct unique and positive identities for themselves–while Apple's goal was to stay in the spotlight by constantly updating their inventions, Xiao Mi's goal was to expand their bases to neighboring countries so that one day they may be known worldwide. These narratives also corresponded with what is known about the two companies outside of the situation in the test, and so it is quite impressive that the students were able to manage their WTs, given the possible rivalries between their companies, in ways that their companies were both represented positively and professionally.

To offer a counter example of professional construction of students' company in WT, let us look in Excerpt 5.12 at Jim's construction of his company's identity in his WT starting from line 110 below.

Excerpt 5.12 *"My company isn't doing very well"*

```
110    Jim:   about (.5) the conference=so,
112           (.3)
114    Jim:   +I work for A-M-D. ↑I .tsk >I am< chief innovation officer,
       Jim:   +GZ→Sus

115           +(.8)
       Pet:   +GZ→Jim, nod

116    Pet:   +innovation?
       Pet:   +raise RH to gesture something
```

```
      Jim:   +GZ shift to Pet, nod lightly

117   Jim:   yeah.=
118   Pet:   =um!
119   Jim:   ↑+I:: manage (.6) everything in relation to innovation
      Jim:    +point down forward

120   Pet:   ↑um:=
121   Jim:   =and:: sometimes (.) I:: also (.) origin-
122          originate my own ideas.
123   Tho:   [um interesting.
124   Pet:   [oh::::
125   Jim:   yea::: h h
126   Joh:   sound great!
127          (.4)
128→  Jim:   and +how abut- (.) +unfortunately↓ (.2) right now::
      Jim:       +GZ→Tho, RH→Tho +GZ shift down

129          my company isn't doing very well.
130   Tho:   ooh=
131   Pet:   =ooh
132   Jim:   we are .tsk >we are at a loss<.

133          +(1.0)
      Tho:   +slow nod
      Pet:   +hold GZ at Jim then nod before shift GZ down

134   Jim:   ↑BUT (.) that's why I really need to- (.5) come to::
135          this conference,
136          (.)

137   Jim:   to meet with +˚you˚ ˚˚people˚˚
      Jim:                 +point BH forward (in Sus direction)

138   Pet:   +oh kay.=
      Pet:   +GZ shift & body turn from Jim→Sus,

139   Jim:   =you are all +.tsk +(.5) very::::: s:kill.
      Jim:                        +small head tilt↑
      Pet:                 +GZ shift back to Jim
140          (.)

141   Pet:   +o:kay.
      Pet:   +shift GZ down

142          +(.5)
      All:   +exchanging nods

143   Tho:   +˚yea:::h˚
      Tho:   +turn to Pet

144   Pet:   +um=
      Pet:   +GZ→Tho

145   Tho:   =+what's [about you?
      Tho:    +open LH to Pet

146   Pet:            [+ah
      Pet:             +GZ down

147   Pet:   yes=ah (.) my company is ah (.) stark industrial.
148   Sus:   [oh:::::]
149   Joh:   [OH:::::] Stark industrial.
```

From the beginning, we can see that he was able to manage his WT quite well, maintaining the audience's understanding organizing his participation without any disruptions throughout. However, starting in line 128 after a potential sequence closing pause, he cut off his nomination of the next speaker and instead launched another expansion to his own WT. His embodied actions also confirmed this interpretation as both his gaze and his right hand were pointing in Thomas's direction at first, but following the cut off, he shifted his gaze down to the floor. What followed was quite uncharacteristic of a professional talk at a trade expo. To account for the purpose of his attending the conference, he cited his company's current business struggle (line 129) and further disclosed more insights of the company's negative financials (line 132).

Such self-deprecated construction of a company's narrative posed some complications for the co-participants who then had to respond to the actions Jim had just projected. We see a long pause in line 133 following Jim's negative narration of his company during which Thomas did a really slow nod while Peter slowly shift his gaze away from Jim. In the end, other members in this group did not come up with any response to Jim at all. When the conversation reached another winding down moment (lines 141–142), all participants just nodded to each other and moved on to the next speaker.

In conclusion, for WT, students need to demonstrate an ability to initiate a WT sequence or transition from the previous topic to WT. The design of their initiation, albeit in the form of a question which afford opportunities for others or a taking the turn for himself, should have a clear audience with a construction which caters to the intended audience. Students should also demonstrate an ability to develop WT topic collaboratively with their peers, showing engagement in each other's WT and initiating expansion sequences when it is relevant. Nominating the next speaker abruptly while their peer's WT is still ongoing would be a major problem. The same goes for failing to address and handle any potential misunderstanding

through initiating and resolving of repairs. Doing so, students should demonstrate that they have sufficient resources and were prepared to cover all the areas required under WT: their position, their company and their job responsibilities. Additionally, students should also display through the design and execution of their WT sufficient understanding of their chosen profession and industry as well as the conference they were attending.

Contact Exchange (CE)

In this section, the sequence organization of how students handle the part of the roleplay that build up to "business card" part of the task requirement will be explored. This exchanging contact activity is often embedded under the WT activity as they brought up their company and work responsibilities in their roleplay conversation, then moved to state possible future collaboration as a reason for exchanging their contact information. The reason why we discuss this activity separately from WT is because of a rather distinct sequence organization of CE talk, which will be discussed below.

From screening through student CE performances in this dataset, there generally were three parts, three smaller accomplishments that formed a whole unit of CE, to the construction this action: (a) topicalizing shared objects between the speaker and his or her recipient, (b) establishing reasons of relations or possible joint projects, and (c) explicitly making a suggestion, a request, or an offer to 'work together' in the future. Afterwards, students may exchange handshakes or name cards before they move on to other business.

From a personal discussion with experienced teachers who were also the original raters for this task, actual business card exchange was not deemed crucial to the organization of CE activity. So, although the descriptor indicated "business card", students who were able to

establish contact exchange appropriately in interaction would not get a lower score just for not having the actual name card prepared.

In the following sections, student competent performances in doing contact exchange will be examined. Then, we will consider also problematic performances and discuss the application of CA in rating these performances.

Competent performance in doing CE. At the core of the contact exchange activity, competent students have multiple ways to execute the focal action of explicitly stating that their respective companies work together. Many are done by issuing an offer, i.e., "we can work together in the future"; some are done by forming a request, i.e., "can we exchange name cards, so I contact you later?".

Before reaching that point in the interaction, what competent students could do that set themselves apart from their peers is that they foreshadowed this contact exchange activity multiple turns before the actual offer or request for contact information took place. In CA, these actions are called pre-sequence (Schegloff, 1988). So instead of saying that students can foreshadow their action trajectory, it can be said that they were able to do pre-request or pre-offer and here, they give reasons for why such contact exchange is necessary, profitable even, in the first place. When a pre-sequence was done successfully, not only were the speakers able to foreshadow the upcoming course of actions that can lead to name card exchange or a commitment for their companies to work together, but they also made the trajectory available for their recipient who will therefore have the opportunities to co-construct their action with the same goal in minds.

In excerpt 5.13 below, Ben issued a pre-offer in line 206 citing his boss's potential interests in Bob's line of work as the reason for contact exchange. Recognizing the trajectory of

the actions, Bob co-completed the sequence as he quickly continued on to suggest that they can 'work together' in the future in line 207.

Excerpt 5.13 *Ben and Bob business contact exchange*

```
204    Ben:   I hear that you: (.5) uh:: make a advertisement in internet=

205    Bob:   =+yes.
       Bob:    +nod

206→   Ben:   oh +yes I think ah (.2) my boss must interest about (.) your job.
       Ben:       +point to Bob

207→   Bob:   yes=maybe in the future: >we can< (.2) +working together.
       Bob:                                          +offer handshake to Ben

208    Ben:   ↑+yes=
       Ben:    +shake Bob's hand

209    Bob:   =↓yes.

210           +(.8)
       Bob:   +hold on to the handshake
       Ben:   +hold on to the handshake
```

Interestingly, Ben's projection of this contact exchange, or Bob's recognition of it, arguably started before his pre-offer in line 206. This activity sequence began with Ben proffering (Schegloff, 2007) the already known information about Bob up for his confirmation in line 204. Students in this dataset sometimes format this turn in the form of confirmation checks just like in this excerpt. The turn was designed to get a go-ahead response from their recipient before they went ahead with the pre-request and subsequently the request for contacts actions. In other words, Ben's turn in line 204 topicalized Bob's work, the object of reference which was later used to build his argument to establish their business contact exchange between the two.

From this episode of interaction, Ben's knowledge of this normative structure and his skills in putting his linguistic resources together to issue the activity are on display. This is evident in the fact that Bob was able to recognize the trajectory of Ben's project by co-completing this action in line 207. Afterwards, we can see that Ben agreed to Bob's offer with a firm 'yes' in line 208 and Bob's sequence closing 'yes' in line 209. The two shook hands over their new agreement (line 210), marking the end of this activity.

In terms of interactional competence, by effortlessly co-constructing this action, both Ben and Bob displayed the knowledge of the sequential order of this action in the context of this test. On the production side, Ben was able to produce a sequence of turns recognizable to his recipient, who on the other hand correctly ascribed Ben's action and fully participated to complete the action with Ben. If raters only pay attention to linguistic accuracy, they might notice only the wrong use of article "a advertisement" in line 204 or the incorrect verb forms following the modals in line 206 and 207.

Assessing interactional competence can highlight students' achievement in constructing actions competently. In this case, Ben has displayed the knowledge of this action sequence and an ability to summon linguistic resources executing required actions which are recognizable to his interlocutor and appropriate in the eyes of the raters according to the "normative standards". More specifically, in this part of their roleplay, Ben displayed that he can pivot the ongoing talk to a sequence where he can exchange contacts with Bob and fulfill one of the requirements of the task. Bob also deserved much credit for excellent recipientship in co-constructing the action. He showed that he recognized Ben's projected action and was able to help bring the action to its completion with more efficiency.

We can compare Ben and Bob's accomplishment in mutually constructing the contact exchange sequence with another pair of student interaction in which one student took on the responsibility to carry out the activity more so unilaterally. Performing the same task, in this excerpt, Ben, again, initiated the establishing contact sequence with Jey as his recipient.

Excerpt 5.14 *Ben and Jey business contact exchange*

```
125→  Ben:   +uh +(.7) +engineering company?
      Ben:       +GZ up +GZ→Jey
      Bob:   +GZ→Ben

126           (.5)

127   Jey:   +yea:s.=
      Jey:   +upward nod
```

```
128→ Ben:   =yes=+↑a:h my- (.) my:: >company is< +ah it-
     Ben:          +GZ up                              +open RH

129          >my boss is interesting about,

130          (.3) setting up the old company here (.) in +Sydney
     Jey:                                                 +nod twice

131          (.4)
132  Bob:    [oh! ]
133→ Ben:    [yes.] so I think (.) ah we can work together
134  Jey:    ↑a:h ↑yea[::h.
```

As this excerpt shows, Ben did a confirmation check in line 125. He issued a pre-offer providing a reason for their future joint venture in line 128 and finally put forward a suggestion for the two companies to work together in line 133. Note that after confirmation check in line 125, the warrant second pair part from Jey came in as a delayed and elongated yes (line 126-127). Displaying his awareness for co-construction in interaction, he waited for Jey to grant a go-ahead signal (line 127) before he put forth his next action citing possible business expansion as a reason to establish the contact between his clothing company and Jey's construction company in line 128 - 130. In contrast to the previous excerpt in which Bob stepped up to co-producing the next action, Jey only supplied a few nods which made her action rather vague. What we would expect to see here is a display of alignment with Ben's statement, a positive assessment token like "that's great!" or some form of acknowledgement token like "oh!" to show that she welcomed the opportunity to establish business contacts between the two. Instead, Jey's nods displayed no alignment, simply a signal for Ben to keep going. She did not show that she recognized the trajectory of this sequence until line 134 after Ben explicitly suggested that they work together in line 133, when we finally see Jey's two-step response: first a change-of-state token (Heritage, 1984a) in an emphatic, high pitched 'ah' and second an agreement response 'yeah,' also done in high pitch.

Interestingly, Bob also participated as a by-stander in this sequence. He provided a verbalized second pair part to Ben's proposal turn in line 132, an action which Jey was expected

to supply. Because Ben's next turn overlapped with Bob's acknowledgement token, it is more likely that Ben did not react to Bob's contribution and just treated Jey's nod as a sufficient recipient action and went ahead with the next action regardless of its ambiguity. In comparison to the previous excerpt, Jey was much more passive in co-constructing this contact exchange action with Ben. The upside is that she did not obstruct the trajectory of Ben's project. However, she did not actively take part, let alone display a recognition of the activity onset in co-constructing this contact exchange with Ben either. In terms of what raters can gauge from her performance as part of a pair co-constructing this action, it is quite clear that her contribution is significantly less substantial than Bob's in the previous example (excerpt 15). This comparison provides an evidential basis for language testers to start formulating a rating scale, for the recipient side of the interaction, which could range from what Bob can do–successfully displaying recognition of and being able to co-construct the activity with his interlocutor–to what Jey can do, which is still co-operative in nature, but takes much fewer interactional skills to execute.

Problematic performances in doing CE. For this particular activity, students who were not as competent in producing this action faced many kinds of challenges that can result in a problematic management of CE. As each of the occurrences of CE was quite unique in what made it acceptable or what made it problematic, what I will attempt to do here is to provide an overview of some typical cases which were found more often in the dataset, in a hope that they would exemplify some notion of standards we come to adopt in our study.

First, one of the most common problems found in this set of roleplay performances is that students misplace the CE action within an overall sequential position of the activity. Excerpt 5.15 shows a case which participants initiated the name card and contact exchange sequence after the role play was already in the closing phase.

Excerpt 5.15 *"Do you want to have my name card?"*

```
224    Ste:   ah so: I think (.4) +my team is calling (.) +so I have to ↑go.
       Ste:                       +RH thumb point to the back +GZ around then at Ben
```

225 (.2)
226 Art: um [˚yea˚

227 Sea: [+okay so [+xx
Ste: +GZ→Ben, nod
Sea: +shift GZ from Ste to Ben
Ben: +shift GZ from Ste to Sea

228 Ste: [+let see]
Ste: +GZ started to shift to Art

229 Art: [I better] +go↑ (.) get going too.
Ste: +lean in, GZ→Art

230 (.4)

231 Ben: [+˚okay˚
Ben: +shift GZ to Ste

232 Ste: [+>˚yah yah˚ yeah.< [x] x=
Ste: +nod, back straighten, GZ→Sea
Ben: +GZ→Ste, nod

233 ?: [let's,]
234 Art: =okay see you,
235 ((all saying 'see you guys' and wave at each other almost at the same time))

236→ Sea: >+do you want to have my name +card?<
Sea: +take out name cards from his pocket
Art: +reach into his pocket

237 Sea: +so: (.) +you can contact me later?
Sea: +hand his name card to Ste
Bro: +reach his pocket

238 Ste: +thank you:::
Art: +hand his name card to Ste
Ste: +receive name card from Sea, GZ down at name card
Buc: +reach his pocket
Ben: +reach his pocket

239→ Bro: +yeah:: [I'm a +C-E-O,
Bro: +RH give out name card to Sea, LH take name card from Sea
+RH give name card to Art
Ste: +taking name card from Art

240 Ste: [thank you::=

241 Bro: =+I'm a C-E-O
Bro: +RH take name card from Art into his pocket

242 Ste: h hh [hh

243 Sea: [+hh+h=
Sea: +GZ→Ste/Bro
Ste: +RH pat Bro in the back

244 Bro: =+˚£I'm a C-E-[O£,˚
Bro: +GZ down on the card in his RH, smile
Sea: +GZ shift to Ben/Buc who're exchanging cards in the background

245 Ste: [+he keeps saying he's a +C-E-O
Ste: +GZ→Sea +LH point at Bro

```
246    Buc:   +˚thank you.˚
       Ben:   +take card from Ben
       Bro:   +hold out card to Sea/Art
       Sea:   +exchange cards with Art

247    Bro:   +call to me?
       Bro:   +move closer, hold out card to Art
       Sea:   +GZ shift to Bro briefly then down


248    Ben:   +and this one's for you guy
       Ben:   +hand card to Sea/Art

249    Buc:   +˚thank you˚
       Buc:   +GZ down, nod

250    Sea:   +yeah,
       Sea:   +GZ at the card Ben's holding out

251    Art:   +good luck +guys (.) +see you la[ter.
       Art:   +RH low wave, GZ→Bro
       Ben:   +GZ up at Art          +GZ down, nod
       Sea:              +take the card from Ben

252    Bro:                                    [good luck,

253    Buc:   +see you.
       Buc:   +GZ→Art
       Sea:   +give card to Art
       Art:   +receive card, GZ→the card

245    Sea:   +good luck. See +you.
       Sea:   +GZ→Ben,Ste,Bro
       Bro:                  +nod,smile
       Ste:                  +nod, GZ down

246    Ben:   +good bye.
       Ben:   +RH high wave

247    ((all back away from group))
```

Initiating the closing sequence, Steve provided a "reason to leave" in line 224 and gained alignment from Arthur and Sean, then other group members to closing the role play in subsequent turns (lines 226-235). (For discussion on the management of roleplay closing, see section 5 below). Note that after the culmination of the good-bye sequence (line 235), Sean initiated a move to exchange name cards with everybody in line 236. This turn was done in a fast pace, but with an accompanying gesture of taking the prepared cards out of his pocket. Other students in the roleplay recognized what he was doing, as we can see Arthur and Sean starting to position their hands to obtain their respective cards in lines 236 and 237. It can be argued that at this point, instead of focusing on the talk–creating opportunity to establish reasons for contact

exchange in interaction–students only focused on the physical exchange of cards as part of the task requirement. This resulted in an unnatural exchange such as between lines 239-244 when Brooke kept on announcing that he was a CEO while comically handing out his cards, an action that Steve, who in this role play was an employee in Brooke's company, oriented to as amusing. Steve went on to give an account for his laughter ("he keeps saying he's a CEO") in line 245.

Based on the literature on conversation closing, activities that take place after a good-bye sequence are what Jefferson (1984a) called a last business action. This position is mostly reserved for reaffirmation of prior arrangements or final thoughts usually to loved ones not present in the conversation. This last business action sequence is not usually extensive because the insertion of this activity puts on hold the conversation closing that was already underway (Schegloff & Sacks, 1973). Unless the students were able to manage a conversation restart, which none of the groups in this dataset did, it should be noted that the action of contact exchange is mostly not effective to be executed after conversation closing had already been initiated and ratified.

Secondly, apart from sequential mishandling of the action, another source of interactional problems can be seen to be rooted in the design of the turn or turns students put together to form their CE. In the Excerpt 5.16 below, Chan initiated a shift to this contact exchange activity in line 252. He did a topic transition by employing "actually" in turn-initial position signaling that a shift of topic (Clift, 2001).

Proceeding with his topic initiation, we can see that Chan did a request pre-sequence in line 252-256 and issued a request for a name card from Frank, his interlocutor, in a latched turn in line 256-257. Sequentially speaking, Chan displayed an awareness that competent speakers do mitigate the imposition of requests by employing a pre-sequence, giving account or justification for the actual request coming ahead. The problem is that the design of his account has a shape of

what is normally employed in the request itself instead of the pre-sequence. So, in the sequential environment where we would expect to see a turn that does a pre-request, Chan cited simply his desire to make business networks instead of issuing an account or a suggestion for what his company and the company of his recipient can do together.

Excerpt 5.16 *"I want to make a network"*

```
251    Ric:   um,=

252    Cha:   =actually I +(.) I- (.3) I want (.6) +I want ah
       Cha:               +raise his RH              +open RH ↺ forward

253           (.) to make a +network.
                            +put RH to his back, upward nod to Fra

254           +(.3)
       Fra:   +start to nod

255    Fra:   [wow.  ]
256    Cha:   [in the] business=can,

257           can you +give me a- (.4) +um a [card?    ]
       Cha:             +RH hold up thumb and finger to about a card size
                                  +BH draw a rectangle in front of him

258    Fra:                                  [+ofcourse.]
       Fra:                                   +reach into his pocket, GZ→the pocket

259→   Fra:   +here's (.) my business +card. (.3) so you can +call me or mail by
       Fra:   +still getting card from his pocket-----------------+hand card to Cha
                                      +GZ→Cha

260    Fra:   +(.3) this ad[dress ]
       Fra:   +point at card in Cha's hand

261    Abi:                [+↑ah  ]
       Abi:                 +GZ→card in Cha's hand

262    Cha:                [+(wow)](xx)+(.6)↑o:kay:.
       Cha:                 +GZ→card,   +nod twice

263           (.4)

264    Cha:   +maybe I will go +to-
       Cha:   +GZ→up at Fra      +RH point forward with card, LH open shirt pocket

265           +(.7)
       Cha:   +GZ→pocket, put card in pocket
       Fra:   +nod, GZ→Cha

266→   Fra:   erhm↑ +Electrolux +has: [many shop] in (.) Sydney.
       Fra:                      +point with RH
       Cha:          +look up, GZ→Fra

267    Cha:                             [(xx)     ]

268    Fra:   maybe you can con-=contact me to about set +your new branch.
       Fra:                                             +point with RH, lean closer to Cha
```

```
269            (.4)
270    Cha:    ↑oh::
```

It might appear that the interaction between Chan and Frank went by quite smoothly. Chan expressed that he wanted to make a network (lines 252-253). Frank gave an acknowledgement nod and a positive assessment 'wow,' an unusual lexical choice for this context, but still issued at a sequentially appropriate position as a go-ahead signal (lines 254-255). Chan issued the request for Frank's name card in line 257, which met with an approval from Frank instantly (line 258) as they chorally completed their turns at the same time.

What is notable in this case is that, in a few turns later, Frank provided a suggestion for how their companies can work together in lines 266-268, a content that was missing from Chan's design of his management of contact exchange done previously. Frank's action can perhaps be viewed as a repair of the action sequence earlier, and by doing so, Frank displayed his familiarity with CE action sequence and was able to achieve being cooperative in supplying what's missing from his friend's action the soonest that he had a chance.

This is a case where teacher raters have to make an informed judgment based on their knowledge of normative occurrences in their intended target domains as well as what students in this population oriented to as normal. In the case of Chan, knowing that he wanted to issue a contact exchange in the form of a request, teachers can use information generated from this roleplay to design a lesson which gives structural guidance on the sequential shapes and content to help Chan or any other students to more effectively execute their actions in the future.

Thirdly, another type of off-target management of CE activity was when students brought up some parts of CE activity but failed to complete the action, resulting in an incomplete CE and a missed opportunity to solidify a partnership with their interlocutors. In this excerpt below, Takeshi initiated the CE sequence with Ben as his recipient. Starting at line 151, Takeshi employed a confirmation check, 'new company?,' as a pivot to transition the topic in their talk

and was successful in securing attention from the group and in particular, from Ben who rightly identified himself as the recipient and provide a go ahead signal for Takeshi.

Excerpt 5.17 *"So… I'm interested."*

```
142    Ben:   +and, .tsk (.4) she ah >she's interest< in (.2)
       Ben:   +RH on his chin, GZ up

144           +>setting up< new (.7) company yeah.
       Ben:   +GZ→Jey

145    Ben:   in Sydney. (.) yes=+SO, (.) I'm arrive
       Ben:                        +GZ up

146           ah almost (.) one week ago
147           (.7)
148    Bob:   Ah[:::

149    Ben:     [+to see new place here.
       Ben:       +put RH forward, GZ→Jey

150→   Tak:   (new) company?
151    Ben:   [↑yeas.  ]
152→   Tak:   [you want] (.) you want event (.) (with)
153           (.4) (with) >grand opening<?
154    Ben:   oh ↓sure.=
155→   Tak:   =if you want I can (.2) I can help you
156           (.3)
157    Tak:   I CAN (.6) make (.7) ah:: (.5) a-

158           +(1.6)
       Bob:   + GZ→Tak, nod, mouthing 'it's okay'

159    Tak:   I can make a (.3) event list
160           (.8)

161    Tak:   event grand opening (.) with uh (nice) +em cee::
       Tak:                                          +open RH

162    Tak:   prit- (.4) pretty +model,
       Tak:                     +open RH

163    Bob:   $phret[ty model$ h[h h        ]
164    Ben:         [↑ah:     hh[h h h      ]
165    Jey:                     [(xx) °model] hhh°
166    Bob:   I like it. H h

167    Bor:   +model ↑ah.
       Bor:   +tilt head, GZ→Tak, cheeky tone

168    Bob:   aw yeah.
169    Bor:   ↑oh.
170           (.3)
171→   Ben:   so I'm (.) interest about it.
172→          can you (.3) give me ah any contract?
173    Tak:   ah (.3) yeah.
174           (.4)
175    Tak:   [This]'s my business card=

176    Ben:   [+oh ]
       Ben:    +receive name card from Tak
```

```
177    Ben:   =+thanks uh yes.
       Ben:    +nod, GZ→ name
```

It is arguable that Takeshi did not treat this second pair part from Ben as a condition before moving on with his next action, establishing a reason for their business connection, in line 152-153 because Ben's go-ahead signal overlapped completely with Takeshi's turn initial. However, we can see Takeshi's orientation to the overlap as he later did a self-repair after a micropause and restarted his turn. Takeshi's attempt at providing justification for a business relation between his and Ben's company was done in a format of a question (line 152-153), articulated in a very simplistic form, yet projecting a clear action to Ben.

After Takeshi secured Ben's attention and brought up possible reasons for their future collaboration, there is a projection of CE completion which can be done via an explicit mention of co-operation of an initiation of name card exchange. At this point, the completion of his CE was still pending, but Takeshi had moved to describe many things his company can provide in his subsequent turns. Until 15 lines later, Takeshi's topic had wound to a stop in line 170. Seeing that the explicit CE request or suggestion was not forthcoming, Ben had moved to issue such turn himself, stating his interests in Takeshi's offer and requesting the name card in lines 171–172. Also orienting to this missing action, Takeshi did a change of state token "ah" before providing his answer "yeah" in line 173.

In summary, among weaker students, interactional problems we observed in this dataset were (a) the misplacement of CE within the overall sequential organization of their roleplay, (b) the off-target composition of their CE actions due to their impoverished interactional resources or poor audience design, and (c) the incomplete execution of CE resulting in a missed opportunity. To reiterate our observations at the beginning of this section on what constituted an effective sequence organization of CE for the *socializing task* among the students in this dataset, competent candidates who were able to successfully manage their CEs were those who could

recognize opportunities to initiate their CE and were able to transition or pivot, an ongoing talk into CE related topic and work together with their interlocutor to jointly reach an agreement to work with one another through an appropriate execution of request, offer, or suggestion. Given the emergent nature of interaction, in each and every episode of CE talk, there was a chance that students may face with misalignment or inactive participation from their intended interlocutors. Students with high level of IC should also display an ability to pursue or re-invoke the CE even in a situation when the sequence progressivity of their CE got sidetracked.

Post-conference Arrangements (PA)

In another recurring activity which all participants consistently performed, students would initiate a plan to do something–a dinner, a sightseeing tour, etc.–together while engaging themselves in a talk about each other's ideas of what to do in Australia besides attending the conference. We will refer to this portion of the roleplay interaction as doing post-conference arrangement or PA for short.

Despite not being specifically spelled out in the original rubric, PA appeared regularly in the dataset. The saliency of PA activity in the roleplay data may be due to many instances in which it was featured both in the learning material for this unit as well as the test preparation guidelines which students also received. Specifically, "talking about future plans" was presented as one of the topics covered in their textbook (See Appendix A), and at the very least all students had practiced making a dialogue asking for each other's plans in one of the exercises. It also appeared on the worksheet that they had to fill out as they were preparing for their roles. Altogether, the topic along with other practices and linguistic expressions became part of the interactional resources available to the students at the time of the test. The omnipresence of PA activity in both the test data and the learning material covered in the classrooms therefore

warranted the study's choice to include this activity as a criterion of this study's assessment constructs.

For a roleplay such as this one, the interaction outcomes of whether a plan was secured by participants during the talk are not the best indicator of an effective PA practice for further extrapolation of students' IC because such outcome was already predetermined by the test activity. Another issue that may hamper an authenticity claim inherent in any roleplay tasks, which is particularly pronounced in PA, was that there could be little or no consequences if students were to make a wildly unrealistic arrangement, unlike a real situation in which accepting an invitation would result in a social obligation to actually do something or spend more time with someone later (for a counter example, see Huth, 2010).

For these reasons, the scores should be awarded based on students' displayed awareness of interactional conventions when it comes to a talk on post-conference plans displayed through their sequential participatory design and execution. To illustrate how students constructed this activity, the analysis below focuses on two kinds of actions--issuing an invitation and making a suggestion–the two main speech acts found as students "talk about their future plans" in their roleplay performances. To determine the varying degrees of success of PA activity, this section is focusing on the extent to which students in each occasion conformed or deviated from normative practices when it comes to producing the two speech acts, and what their action compositions revealed about their IC in doing PA.

In the first three examples, their PAs were organized in a series of inquiries (Schegloff, 1986) into each of their plans after the conference ended. Each student would take a turn to talk about their own plans. Once they had completed their contributions, the current speaker would select the next speaker using turn allocation devices like "what about you" to connect their turns together in one larger PA sequence.

Within this organization of PA, making a joint activity arrangement then would be a local occasion which spontaneously generated (Drew, 2018) when two or more students had established that their plans overlapped, affording a ground for an invitation or a suggestion for a joint trip to take place. Excerpt 5.18 shows a portion of talk when Amber issued a suggestion as the prior talk about plans activity provided an opportunity for her to do so. Before this excerpt, Amber had earlier shared her plan to visit the Sydney opera house with her brother after the conference. In line 373, she nominated Mark to take the next turn, in which he took it up and issued an announcement that he, too, had planned to visit the opera house.

Excerpt 5.18 *"We can go together!"*

```
373    Amb:   .tsk (.) How about you.
374→   Mar:   I think I'm:: (.) going to (.) >Opera House<,
375           same with you.
376→   Amb:   [O:::h!
377    Tim:   [wow!!
378→   Amb:   yeah then (.) we can go together.
379           (.2)
380    Amb:   We will have time to (.4) grab some dinner first=
381           =>and then we can go< (.) to (.) the opera house.
382           (.5)
383→   Amb:   yeah?
384    Tim:   Ye::[:ah!
385    Mar:       [yeah.
386    Amb:   yeah [should-
387→   Lee:        [Let's go to dinner.
388           (.2)
389    Amb:   OKAY↑ [yeah then] (.) we can (.3) ah [swing our] (.) own way.
390    Tim:         [↑oh yeah.]                    [I::      ]
```

Mark sequentially organized his PA in two parts. First, it was the announcement of his plan in line 374 which was constructed as a spontaneous decision, as evident in the "*I think*" turn initial, the prolonged "*I'm*" embodying his thinking action and the rising intonation at the end of the turn. Second, he built on what Amber had shared earlier and commented in line 375 that their plans were to go to the same place. This move from Mark made explicit the ground needed for either of them to launch the PA activity. It was quite crucial at this point to note that Mark's turn design had made it possible for Amber to initiate PA in the next turn, which she did. So, after her

change-of-state token in line 376, Amber issued an explicit suggestion for them to go together (line 378).

Amber's suggestion was done quite seamlessly as she was able to connect her suggestion to the prior topic of talk by starting her turn with "*yeah then*". In a normal situation, following a suggestion in the first pair part, the next projectable action in response to the suggestion would either be an acceptance or a declination. These are non-asymmetric possibilities which tip in favor of an acceptance (Pomerantz, 1984). The delay that followed Amber's suggestion in line 379 was hearable as a hesitation which could be indicative of a declination. Amber showed an orientation to such convention as she quickly pursued a relevant response by giving more detailed information regarding her proposal in lines 380-381. Following her second attempt in pursuant of a response, Amber then issued a question designed to elicit positive response "*yeah?*" in line 383, to which Tim and Mark finally responded with acceptance in lines 384 and 385.

It was notable that Mark's participation deviated considerably from what could normatively be expected. Not only that he did not provide a timely response, resulting in two delays in line 379 and 382, after his acceptance in line 385, but he also did not take an active role in the talk about further arrangements at all. On the other hand, other students, Tim and Lee, appeared to have treated themselves as legitimate audience also of Amber's proposal. Showing his alignment with the ongoing talk, Lee even echoed part of Amber's suggestion to go get dinner (line 387). The sequence was brought to a close when Amber and Tim re-iterated their commitment to the plan and talking about how they would later disband after (line 389) put a finality to their discussion and in effect closed down their PA sequence.

The next excerpt shows another occasion of spontaneous initiation of PA activity from a series of inquiries about each other's plans after the conference. Excerpt 5.19 shows Tara's

execution of his PA unilaterally as an announcement instead of issuing an invitation or a suggestion.

Excerpt 5.19 *"I'm with him"*

```
525    Pan:   +how about you guys do you have plans to go +anywhere?
       Pan:   +GZ→Tar&Tan, RH→Tar&Tan                        +GZ→Sut, RH→Sut

526           (.8)
527    Sut:   um:::::
528    Tan:   I think after: >complete this conference< I (.) will go uh::
529           (.6) Sydney opera house.
530    Pan:   [ah::: Sydney o-
531    Tho:   [oh : : : : : : : :
532    Ath:   [oh : : : : : : : :
533           (1.2)
534    Tho:   we[ll (xx)
535    Tan:     [that's famous (.3) uh famous uh place
536           (.4)
537    Tho:   (x x) (.) (got) a sightseeing spot.
538           >everyone has to go there<.
539    Tho:   what about +(.) you?
       Tho:               +RH point to Tar

540           (.4)

541→   Tar:   I (am) (.) I will come with +(.6) him=
       Tar:                                +RH place on Tan's back

542    Tan:   =he's with me yes.
543    Pan:   [ah::
544    Tho:   [oh::
```

After Athena had talked about her plan, Panu turned to Tara, Tan and Sutham to inquire about their plans. While Sutham was being hesitant, Tan then took the turn and shared his plan that he might go to Sydney opera house afterward. Other group members oriented to Tan's incomplete turn as everyone waited in silence (line 533) after their acknowledgement tokens of Tan's opera house announcement. After the long pause, Tan reluctantly oriented to the silence as signaling that his turn was incomplete and provided what appeared to be an account for his choice to go to the opera house ("*that's a famous place*") in line 535. Another pause in line 536 was readable as another indication for the group's orientation to Tan's incomplete sequence of actions. But, apart from his account for his choice, what seemed to be missing from Tan's turn was for him to explicitly mark that he had said everything he needed to say and close down his sequence by way of selecting the next speaker. Thor's contribution in the next turn (lines 537-

539) partly confirmed our analysis as, from what was hearable, he quickly restated the reason why everyone should go see the opera house before appointing Tara as the next speaker.

Unexpectedly, in line 541, Tara issued an announcement that he and Tan would go to the opera house together. The design of his turn strongly implied that the plan between Tan and himself had been agreed upon prior to this point as he also placed his right hand onto Tan's back in accompanying his turn, a gesture which might have been designed to show closeness or friendship between Tan and himself. It came as a surprise to Panu and Thor, both of whom treated this new information as a surprise, evident in their change-of-state tokens in lines 543-544. This is because in Tan's earlier turn when he shared that he might go to the opera house after the conference, Tan did not indicate that Tara was also part of his plan. Also, granted that the previous interaction in the roleplay provided a context and history in which they were then shared, based on how the roleplay unfolded so far, Tan and Tara had just introduced themselves at the beginning of the roleplay and there was no talk about the two going to the opera house together in roleplay up until that point. Tara's PA design was highly problematic because it rendered his action interpretable as a blunt assertion of himself into Tan's plan without any prior talk or arrangement. Still, what was curious about the roleplay interaction was that Tan then aligned with Tara's assertion in his action completely (line 542). It could be that the students who played the roles of Tan and Tara had agreed upon the arrangement before starting their roleplay. However, this is unknown to the analysis given that we do not have access to the students' discussion during their preparation. This will for now remain an unsubstantiated speculation. Nevertheless, the quickness of Tan's response seemed to point to a rather troubling preference shared by some participants in making an agreement and getting the desired outcome over interacting genuinely in the roleplay to display their interactional competence.

Excerpt 5.20 shows a continued talk from the previous excerpt to display Panu's initiation of his PA through self-invitation.

Excerpt 5.20 *"I have a plan to go there too"*

```
545→  Pan:   actually I have a plan to go (.) there too=
546          =can I:: (.) accommodate you?

547          +(1.1)
      Tar:   +GZ turn to Tan

548   Tan:   yes yes.
```

Here, Panu's formulation of PA activity was designed as a self-invitation issued in the form of a request ("*can I accommodate you?*" in line 546). Panu also provided a preliminary account for his request in line 545, which he packaged with a clause-initial "actually" which mark the upcoming information as new and a "too" at clause final position to connect his utterance to the ongoing topic.

Raters can then consider the design of Panu's request in line 546 to determine its effectiveness. Evidently, there was some hesitation which resulted in a significant delay following Panu's request. It could be that Tan and Tara treated the self-invitation from Panu as difficult to answer to because their response might have been a dispreferred one. Or, it could be that Tan and Tara both relied on the other to take the turn resulting in a turn evasion in line 547. Their eventual "*yes yes*" answer in line 548 seemed to favor the latter analysis.

The three examples above offered cases of PA interaction episodes at the same sequence location so that we can compare different ways participants designed their turns to create an opportunity to build their PAs. What each occasion had in common was the characterization of the event being an emergent one–meaning that it was created locally as the talk progressed–as opposed to an already existing event which we will get into later. In Excerpt 5.18, an announcement of a common activity warranted the other participant to issue a suggestion for a joint trip. In Excerpt 5.19, we see a turn design which forwent the invitation altogether, creating

a contingency for other participants to handle such unexpected announcement afterwards. In Excerpt 5.20, we see how instead of simply issuing an announcement of a coinciding plan, participants can design their turn to use it as an account for self-invitation and thus giving themselves an opportunity to initiate the PA.

For this, we recognize the interactional competence required for students to make relevant the PA activity in interaction that was both recognizable to raters and preferably also the participants in the roleplay. Therefore, in this activity, student performances which demonstrated a higher level of IC would be the ones which appropriately initiate PA activity or transition from ongoing talk such as Amber's timely suggestion for a joint excursion which touched off of the ongoing talk in Excerpt 5.18 or Panu's use of a common plan as an account preliminary to his self-invitation in Excerpt 5.20.

When students did not formulate their PA sequence as an emergent occasion, we can see the post-conference activity being characterized in talk as something which had been prearranged outside of the occurrence of the roleplay. In this formulation, the mere mentioning of the activity can be hearable as a pre-invitation. This sequence organization of PA initiation resembles what Drew (1984) described as an invitation sequence in that the act of reporting can project or elicit proposal or arrangement from the recipient, allowing the speaker to avoid making an explicit proposal that may be rejected.

In Excerpt 5.21, we will consider two examples from the dataset showing how students could formulate a statement about a place as a pre-invitation.

Excerpt 5.21 *"If you don't hesitate you can join me"*

```
309    Ben:   +ah (.) if you don't know there-
       Ben:   +GZ→Bob

310           there is a (.) uh (Medusa) Greek hotel=
311           =which is the (.) one of the most famous restaurant in Australia.

312    Bor:   +medu[sa ]
       Ben:   +GZ shift to Bor

313    Bob:         [huh] medu[sa ]
```

```
314    Ben:                  [YES].uh it's very (.) famous yes.
315    Bor:   °(really)° oh::

316→   Ben:   if you uh don't hesitate you can +join me
       Ben:                                  +RH point to himself

317            (.3)
318    Ben:   I'm and my boss will go there °tonight°

319    Bor:   oh [tonight
320    Tak:      [tonight!=
321    Ben:   =[yeah.
322    Bob:    [tonight
323            (.8)

324    Bob:   HEY (.) +hhhh .h +I want to go with you↑=[Hhh
       Bob:           +GZ→Jey   +GZ turn to Ben

325    Ben:                                          [oh it's o[kay

326    Bob:                                                    [+I want (x)
       Bor:                                                     +GZ,point→Tak

328    Bob:   +[everyone
       Bob:   +GZ,point→Jey, then shift GZ to Bor
```

Ben mentioned a restaurant he claimed was the most famous in Australia (lines 310-311). Other participants reacted with a surprise over the restaurant's name, creating an insertion sequence between lines 312-315, in which Ben responded with a strong confirmation of his claim over the restaurant's fame to restate his earlier statement. After an acknowledgement token from Bor, Ben issued his invitation explicitly in line 316 using a conditional sentence structure, one of the commonly used forms to issue invitations (Drew, 2018), showing his functional syntactic resources for mobilizing invitations. His used of preliminaries or pre-invitation was creditable, even though it was not quite target-like. His formulation of pre-invitation was not quite recognizable as a pre-invitation because it only reported the existence of this restaurant rather than providing information about the prearranged dinner, which he did instead in line 318 after his issued invitation.

Excerpt 5.22 is taken from the same roleplay performance. Focusing on the interaction between Jey and Bob, we can see that Bob treated Jey's report of a great restaurant near her hotel

(line 292) as a possible preliminary for initiating a PA sequence, which he formulated later in line 304 in the form of a self-invitation.

Excerpt 5.22 *"Can I go with you?"*

```
292    Jey:   there is a greats (.6) +restaurant is +near my hotel.
       Jey:                          +BH open        +BH move to her right side

293    Bor:   [hotel,]
294    Bob:   [A h ::][:!
295    Jey:          [yes.
296    Bob:   really?

297    Jey:   +˚yeah˚.
       Jey:   +nod

298           (1.6)
299    Bob:   you you go there (.) this dinner?
300           (1.1)
301    Bob:   this dinner you want to go,
302           (.2)
303    Jey:   ˚ye:s˚.
304    Bob:   can I go with you?

305           +(1.7)
       Bob:   +GZ→Jey, then tilt head slowly, then nod twice to signal Jey

306    Jey:   +oh! +yes.
       Jey:   +upward nod
                   +nod fast several times

307    Bob:   aHhh [>+youwouldliketojoin< us?]
       Bob:         +GZ shift to Ben

308    Jey:        [+you can go with me       ]
       Jey:         +GZ→Bob
```

Interestingly, it appeared that Jey did not plan for any next action after she reported about the restaurant in line 292. There was a clear break of more than one second long pause (line 298) after Bor and Bob, each individually engaged in a repair sequence with Jey to confirm her report of the restaurant (lines 293 and 296), where Jey's potential invitation was hearably missing. So, although the sequence and design of Bob's self-invitation in line 304 was marred with awkwardly placed pauses and action design, it still displayed his awareness of one possible sequence organization of PA actions. His question in line 299 and 301 pursued the trajectory of Jey's earlier restaurant statement as a pre-invitation. After Jey's confirmation, which he treated as a go-ahead signal for his next action, he then issued his request to make a PA between himself and Jey.

At this point, Jey's inadequate competency in managing her PA activity had become more recognizable. Her slow uptake on the activity being conducted between herself and Bob resulted in a stand-still for almost two seconds (line 305) after Bob's explicit request. During that second, Bob displayed a range of embodied actions. He changed the angle of his head, slowly turning it sideways when the sequence had reached a point where it became apparent that Jey's next turn was not forthcoming. He then twice nodded to Jey, a move which may have been designed to prompt her to take the turn or to signal that she should accept his request. What we see next was that Bob's head nod resulted in Jey's change of state token at the beginning of line 306, which she issued with an upward nod, followed by an alignment token "*yes*" which was accompanied by many nods. She provided a second pair part to Bob's self-invitation in line 304 two lines later. It would be considered a delayed response, but neither party cared to account for the delay. In fact, Bob quickly moved on from the PA project between himself and Jey. He designed the next turn as a closing sequence in the third position in line 307, offering what might be an acknowledgement response to Jey's action before immediately turning to invite Ben to join them for the dinner, an event for which he did not quite have the authority to issue an invitation, in the same turn. Jey's turn in line 308 was an elaboration to her alignment token in line 306 which came in an overlap with Bob's invitation in line 307, rendering her turn redundant to the overall organization of the interaction. At the same time, Bob's minimal closing sequence can also be considered premature and abrupt and thus also visibly problematic.

From the above examples, we can see that organizing PA posed many interactional challenges for students to handle in the moments of performing this activity with their peers. What was required at each moment was highly contingent upon the design of the previous turn. Depending on how they characterized the joint event, be it a spontaneous locally conceived occasion or an already established event existing prior to the talk, students need to initiate PA

ideas, build up an invitation or suggestion to appeal others, and display alignment in their response to others' proposals. Students with lower IC for this activity were often seen to struggle in recognizing the onset of PA activity (like Jey in Excerpt 5.22 or Tan in Excerpt 5.20) and failed to display a timely uptake of the invitation being initiated. Even among students who displayed a recognition of PA initiation sequence, which in itself made a huge difference in the progression of the activity, the execution of their turn designs can still be seen as a little awkward (like Ben in Excerpt 5.21 and Bob in Excerpt 5.22) due to their turn designs.

What constantly appeared in the data was that students treated the PA as completed when they had received a confirmation from their co-participant for their PA proposal. It was rare to see interactional work past such point (although see an example of one such exception in excerpt 20) to pursue the specific details such as time and meeting location for their later gatherings. Future design of learning material can address this point, so that students can naturally progress from seeking an agreement to joint activity to discuss the specifics of the meeting arrangements.

So, in concluding how students managed and displayed their ICs during the PA activity, from our current observation, stronger candidates are those who display a keen awareness of sequence organization of invitations or suggestions and are able to initiate them in sequentially appropriate manners. Stronger candidates also showed an ability to promptly display alignment when any post-conference proposals were issued and to provide sufficient account for their actions, either a preferred one or dispreferred one. Finally, stronger candidates were able to design their actions in accordance to the ongoing progressivity of the group's PA talk. As the interaction unfolded, one challenge facing the students was that they had to account for the accumulating arrangements. One participant may have agreed to more than one post-conference activity given that the times were not overlapping.

Activity termination (AT)

Last but not least, before all roleplays could come to an end, students had to co-ordinate their efforts to bring forth activity termination (AT) or the conversation closing sequence to the group interaction. Exiting a conversation requires interactional coordination from participants, as Schegloff and Sacks (1973) noted: simply to stop talking is not a solution to the closing problem in interaction (p. 295). The familial set of expressions often used in terminal exchanges such as "*good-bye*", "*see you*", etc. has been featured time and again in conversation lessons and textbooks. However, to display their IC in doing this activity, it is important that students were able to employ such expressions in a sequentially appropriate placement and to design their actions in a way which recognizably fits the target expectation of AT performance.

Conversation-closing sequence can be properly initiated with a 'pre-closing' when it is placed at the end of a topic (Schegloff & Sacks, 1973). In this situation, pre-closings can function to signal a possible occurrence of conversation closing in the next action. Most students in this dataset displayed an awareness to the pre-closing move foreshadowing their actual closing sequence; however, there was a variation of the quality of their executions, which we take to reflect the levels of their IC in doing conversation closing. Typical cases of how the students managed their pre-closings in AT are shown below in excerpts 5.23-5.25. In all the examples chosen to show how students employed their pre-closing initiations, we wish to point out some considerations over the quality of their executions in our analysis below.

Excerpt 5.23 *Pre-closing and closing sequences by Bor and Ben*

```
360          +(3.3)
       Bob:  +during this pause, exchange GZ with Bor, nod
       Bor:  +during this pause, exchange GZ with Bob, nod

361→   Bor:  [+let's go?]
       Bor:   +GZ→Tak

362→   Ben:  [+okay     ] I think (.) u::m it's about time that it +was back.
       Ben:   +GZ shift to Bor                                    +point to the back

363          I think we should go in no:w.=
364    Bor:  =okay::
```

In the roleplay in excerpt 5.23, we saw the pre-closing initiated by Bor and Ben (lines 361, 362) after a closing of what they treated as the last topic. There was a sizable moment of silence in line 360, during which Bob and Bor silently sorted out the next speaker to take the turn and possibly came to an agreement on the next relevant action. Bor's pre-closing "*let's go?*" occurred in an overlap with Ben's pre-closing turn initial "*okay*". Ben proceeded to produce his pre-closing in full, giving an announcement that would warrant for conversation closing for the group in line 362 ("*it's about time … we should go in now*"). While Ben's pre-closing composition provided an account for the impending conversation closing, the same cannot be said about Bor's pre-closing composition as his action merely suggested that the conversation could end and invited alignment from others to help complete his closing initiation. For this reason, for our assessment of students' IC in managing AT, at least at this pre-closing stage, a composition which includes a token of closing-warrant announcement would be rated higher than a warrant-less statement which according to Schegloff and Sacks (1973) were seen further down the sequence progression after the closing has already been achieved.

Excerpt 5.24 shows a case in which a pre-closing is designed to only excused himself instead of a "*we should all go somewhere*" like the one we have seen in Ben's pre-closing composition. This type of "*I gotta go*" statement is also a common technique for closing down a conversation.

Excerpt 5.24 *Pre-closing and closing sequences by Jack and Kim*

```
278→  Jac:   uhah:: (.) I think:::↑  I::=there's many
279           (.3)
280   Jac:   ah:: customer f- (.) for my company.
281           (.5)
281   Jac:   and I think::↑ (.) >after that I will< (.5) go to:::
282           (.6)
283   Jac:   uh::: explain how (.5) my (.) system work,
284           (.5)
286   Mat:   oh:

287           +(1.0)
      Kim:   +nod, GZ→Jac then shift to Mat
      Mei:   +nod

288   Jac:   (guest) (.) and
```

```
289          (.6)

290    Jac:  +it's
       Jac:  +raise his RH, GZ→his wrist then GZ up to Kim

291          +(2.2)
       Kim:  +check his watch, gaze at Jac then give him a nod
       Jac:  +GZ→Kim's watch

292    Jac:  ah::

293          +(.5)
       Jac:  +GZ→Kim, smile

294    Jac:  h::
295          (1.4)
296    Kim:  ah

297          +(.9)
       Kim:  +GZ→his watch

298    Kim:  +Ah ye- yes
       Kim:  +continue GZ→his watch

299          +(.4)
       Kim:  +GZ→Jac

300→   Jac:  +ah so::: ah I (.3) have to +go.
       Jac:  +GZ→Mei                      +nod


301          +(1.9)
       Mei:  +nod repeatedly
       Kim:  +check his watch

302→   Kim:  ah yes so (.) ah (.3) I have to go now
303          I have car test at

304          +(.9)
       Kim:  +shrug shoulder, open palm LH, smile

305    Mat:  +£Okay£
       Mat:  +GZ→Kim
```

However, one may note that Jack's composition of his pre-closing was not quite proficiently executed. He initiated an action in line 278 with a statement about what had been going on at his exhibition ("*there are many customers… I will go explain how my system works*". He then seemed to struggle with completing his pre-closing, leaving multiple lengthy unaccounted pauses and many attempted word-searches before finally formulated the pre-closing token "I have to go" in line 300.

While raters can award him for his displayed knowledge of sequence organization of AT, he can still improve on mobilizing necessary resources to execute his action more effectively.

Also, we see in this excerpt that Jack's co-participants, Mei and Kim, had started to display their alignments with his pre-closing move since line 287 during that one second pause in between Jack's ongoing production in which Mei issued a nod directed at Jack and Kim visibly checking his watch for the time. Jack seemed to be focusing solely on his own production of pre-closing such that he missed opportunities to adjust his action to reflect others' contribution as his pre-closing unfolded.

In the last example showing how the students managed their pre-closing in their AT section, Excerpt 5.25 displays a severely disruptive pre-closing initiation by Thor.

Excerpt 5.25 *Pre-closing and closing sequences by Thor*

```
561    Sut:   becuz I like to surf.
562           and you want to go there as well for surfing.
563    Ath:   I love sunny more than uh (.) winter.
564           (.4)
565    Ath:   I am (.) got a cold.
566           (1.6)
567    Sut:   cause you're from Europe
568           and that's [(.)  ]raining [there.
569    Ath:              [↑yeah]

570→   Tho:                              [+OOH!
       Tho:                               +look at his watch

571    Ath:   it's one [of the  ] big problem.
572→   Tho:            [OH NO n-]
573           (.2)
574→   Tho:   WE'RE OUT O'TIME THE CONFERENCE IS ABOUT TO START!
575           we m- [we must] ↑hurry.

576    Ath:          [+↑oh!  ]
       Ath:           +look at her watch
```

Before the starting point of Excerpt 5.25, Sutham and Athena had both expressed their plans to visit a nearby beach. In between lines 561-571, we can see that they were talking about the reasons they had chosen to go there. Thor's pre-closing was initiated in line 574 in a shouting voice in an announcement that time was about to run out and that they all had to end their conversation soon. The method by which Thor secured a turn to initiate his pre-closing was quite disruptive. In line 570, he loudly interjected "oh!" which overlapped with Sutham's ongoing turn in line 568, and issued another negatively implicated exclamation "oh no!" in line 572 which, again, overlapped with Athena's turn in line 571.

In comparing different methods that students employed in formulating their pre-closing sequence, the key distinctions between different levels of execution quality can be observed in their placements and compositions of the pre-closing turn. Students with a higher level of competency, such as Ben in Excerpt 5.25, are those who can initiate the pre-closing sequence after the previous topic had come to a close and are able to design their turn to effectively provide "reason for leaving" that is coherent with the ongoing context of their interaction. Students with a lower level of competency, such as Jack in Excerpt 5.24, can display an awareness of appropriate placement of pre-closing sequence, but may struggle with the design of their action, resulting in a less efficient, less recognizable way of pre-closing execution. A more serious problem is when their sequence initiation was disruptive to the ongoing talk, like in the case of Thor in Excerpt 5.27. Disruptive initiation of pre-closing sequence is more troublesome because, first, being disruptive reflects problems in the control of their turn-taking management (Cekaite, 2007) and, second, it may interrupt other participants' performances on previous interactional activities.

When the pre-closing got taken up and ratified by their co-participants, the conversation closing sequence would progress to the actions Schegloff and Sacks (1973) referred to as terminal exchanges, which constitute the last sequence of talk required for all involved parties to collaboratively close down the roleplay. Students in this dataset tended not to have much difficulty with doing the terminal exchanges. In Excerpt 5.26 below, we can see that students exchanging "*nice to meet you*" tokens at the end of their roleplays, thus achieving a proper close to their AT activity.

Excerpt 5.26 *Matt's last business and terminal exchanges*

```
305    Mat:   +£Okay£ +(.3) I'm getting a drink +you want to join me?
       Mat:   +GZ→Kim +GZ shift to Mei              +RH→Mei, then to himself

306           (.5)
307    Mei:   oh: I have to go to (.) [airport
308    Mat:                           [oh
309    Mei:   hh [sorry
```

```
310    Mat:      [alright
311    Bam:   (xx) I can +join you=
       Bam:               +open RH to Mat

312    Mat:   =+yeh yeah (.2) come!
       Mat:    +GZ→Bam

313           (.3)
314    Mat:   Nice to meet you guys,=
315    Kim:   =nice to meet you.
```

Closing sequence is a routine which once initiated would strongly project the upcoming closing to the session of talk. Minimal last business sequences which can occur after the closing initiation include exchanging of minimal affiliative tokens or a reinvocation of previously made arrangements (Button & Lee, 1987; Schegloff & Sacks, 1973). However, engaging in extensive last business talk which would in effect constitute a reopening of a topical talk would require participants to do some interactional work to divert the course of conversation away from its impending closure. We have seen one example earlier in Excerpt 5.15 when we discussed the placement of contact exchange (CE) activity in the overall sequential organization of the roleplay. While it may be okay to bring up name card exchange as part of a reinvocation of the previous arrangement, initiating the whole CE sequence after the closing section has been initiated and ratified by other participants would not be an effective management of neither their CE nor AT.

In conclusion, in managing their task termination or AT, students have to display an awareness of pre-closing and closing sequence organizations. The placement of AT sequence should be initiated after a closure of the last topic of their conversation rather than disruptively brought up while other topics or actions were still ongoing. In designing their pre-closing initiation, one key feature which we take to be an indicator of their IC is that the quality of their "reason for leaving" which served as a warrant for the conversation closing. During the closing sequence, students may bring up some last business sequences such as a reinvocation of prior

arrangement or any minimal affiliative work before following through with terminal exchanges. However, their last business talks should not be hearable as extensive or constituting a new topic.

Recipient Actions

One of this study's aims in exploring the construct of IC for language assessment is to expand the focus to encompass the quality of students' management of listenership, or in CA term, the social actions which participants accomplished in response to a range of firsts or sequence-initial actions (Schegloff, 1968; Schegloff & Sacks, 1973). Beyond the notion of recipient action simply as display listenership, CA studies have covered many topics under recipient actions which documented social action accomplished in response positions. Prominent studies included Heritage's (1984a) analysis of "*oh*" in which he showed that in certain interactional environment it can be used to display the change in recipient's knowledge following what has been said in the prior turn. Jefferson (1984b) showed how laughing together is a way in which recipients can display alignment and affiliation with the prior speaker. From a minimal response token to a more extensive answer to a question or even a mere nod during an episode of storytelling, these tokens carry out important social actions and thus form a crucial aspect of IC.

For this roleplay *socializing task*, there are three recipient actions which we can highlight as part of students' interactional accomplishments: (a) their ability to display understanding (intersubjectivity), (b) their ability to display alignment (action orientation), and (c) their ability to display affiliation (affective stance). These three aspects of the students' recipient actions are by no means exclusive categories. For example, a continuer '*uh huh*' in the context of storytelling can be taken as a display of both understanding and alignment (Schegloff, 1982b). The same token '*uh huh*' response in the context of joke-telling can be taken as a disalignment as

well as a display of disaffiliation. For recipient actions, these dimensions can be seen as tightly interrelated; however, they each play different roles in shaping recipient participations and variably become relevant in different circumstances depending on the sequential placement of the recipiency display and the kinds of actions to which they are in response (Steensig, 2012).

In designing our new rubric for the purpose of assessing these different aspects of IC in recipient actions, this study is treating these aspects as different interactional constructs. To this end, we will discuss, in turn, each of the recipient actions (understanding, alignment, and affiliation) and present typical cases when deviations from the expected recipient actions occurred in the dataset.

Understanding (U). Displays and negotiations of understanding form the very foundation for social interaction. Through the design of each turn at talk, participants display their understanding of the prior turn and deal with what they deem as troubles along the way. Success in doing so allow for the participants at talk to maintain their intersubjectivity.

For L2 speakers, this construct is particularly useful as it augments a rather mainstay concept of intelligibility as a standard for L2 spoken discourse. As for intelligibility, the focus is solely placed on the language production, while the notion of intersubjectivity also embraces the process in which mutual understanding is negotiated and displayed. On the productive side, action initiators can monitor the audience's understanding and, if necessary, modify their utterances through the use of repairs. The audience is equally responsible in displaying their understanding of prior utterances. The repair toolkit available to the first speakers is also available for the recipients should they need to address any understanding gap between both parties. For this reason, displaying understanding is a highly dynamic construct as both the action initiators and their counterparts constantly display and maintaining their understanding on a moment-to-moment basis.

In this roleplay task, there were few occasions in which one of the students in the roleplay failed to maintain intersubjectivity with their co-participant. This included only cases in which misunderstanding or potential misunderstanding did not get resolved even though one or more party had attempted to initiate a repair.

A case which illustrates this point was presented previously in Excerpt 5.10, a talk taken place during Ann's WT episode. A part of the talk which is relevant to our current discussion is reproduced below in Excerpt 5.27.

Excerpt 5.27 *Ann's failure to provide a repair solution*

```
192    Ann:   um (.5) my job is (.) about create and develop spaceship,
193    Chr:   ↑um[:: ]
194    Ann:      [and] (.) also include about working about
195           sound ↑program.
196→   Hug:   sound program?
197→   Chr:   ˚sound problem?˚
198    Ann:   we make (.2) astronaut safe from ↑danger.
199           when they ↑working.
200    Kai:   oh!=
201    Ann:   =that's [my job.
202    Hug:           [↑um.
```

In line 192, after spending half a second to summon her thoughts, she started talking about her job responsibilities which continued on in lines 194-195. After she mentioned one of her responsibilities with "sound program", her co-participants, Hugo and Christian problematized the item. Hugo reiterated the item with rising intonation in line 196 and Christian, in a quiet voice, muttered "sound problem" also with a rising intonation in line 197. To the analyst, while both Hugo and Christian were both initiating a repair for "sound program", they seemed to have oriented to different sources of troubles associated with it. Christian designed his repair initiation as a hearing problem. For Hugo, there may have been some issues he had with the meaning of "sound program" and its rather obscure connection with her job making spaceships. Instead of taking up Hugo and Christian's repair initiation and providing any solutions, Ann's next turn proceeded as though no repair request was initiated. She offered the last piece of information about her job responsibilities (lines 198–199) and indicated that she has

completed her part in line 201 with “that’s my job”. Failing to monitor their co-participants’ understanding can reveal the students’ orientation of a certain portion of the roleplay more like a monologue than a dialogue. These included cases of failure to respond to repair initiations, such as Ann’s case as illustrated above.

For students in this study, a misunderstanding could reflect problems beyond an understanding at a surface level meaning. Misunderstanding can sometimes reflect a failure to recognize the projected action in the prior action or a failure to match its affiliative stance. These problems reflected poor alignments and a problematic display of affiliation, the two recipient actions which will be discussed below.

Alignment (AL). According to Steensig (2012), aligning responses accept the action projections set out by the first pair part. These projections include the kinds of activity, the proposed interactional roles, any presuppositions assumed by the speaker, and also the designed preference. For example, when you respond to a question “*can you pass the salt?*” by handing that person a shaker of salt, you aligned to the action as a request, instead of simply an inquiry about your ability. Since the former is the more conventional interpretation of the utterance, you therefore showed your ability to properly display alignment by passing the salt.

Although it is worth noting that an alignment display can also be taken as a display of understanding, a display of alignment reflects understanding beyond surface meanings. As we have seen in the above example when we consider a request of salt, to echo Stivers, Mondada, and Steensig (2011), an aligning response provides a display of “structural level of cooperation” (p.20). In other words, it shows recognition beyond literal meaning, including a recognition of the activity in progress and the ability to design a response in relation to such projection.

Interestingly, the group format of this roleplay task sometimes afforded raters direct comparative cases of different alignment work done by multiple students in the same sequence.

For example, in Excerpt 5.28, which shows a portion of talk following the one in Excerpt 5.6, Takeshi, the conference organizer, asked other participants about how the conference had been for them in line 71. Both Bob and Ben each provided their responses to Ben's question in lines 73 and 76 consecutively.

Excerpt 5.28 *Comparing Bob and Ben's alignment displays*

```
71     Tak:   um how (.) how about this expo?
72            (.4)

73     Bob:   +.tsk (.) I think it's +$cold today$ hh [hhh]hh
       Bob:   +GZ turn up              +LH rub upper right arm

74     Tak:                                                [ah↑]
75     Ben:                                                [hhh]
76     Ben:   It's very nice
77            (.4)

78     Bob:   but it's +↑good↑.f
       Bob:            +nod several times
```

Given the role of Takeshi being an organizing staff who had just introduced himself to a group of conference participants, his question in line 71, "*how about this expo?",* was designed be recognized as him doing his job to make sure everything was going smoothly at the conference. Bob's response about the weather in line 73 then can be interpreted as a misalignment as it did not respond to such projection interpretable from the sequential placement of Takeshi's question. Ben's response, in line 76 on the other hand showed his recognition of Takeshi's prior action evident in the design of his response "*It's very nice*", which was hearable as a comment about the conference's organization being "*very nice*".

Outside of the opportunities such as this one, it can be challenging for raters to inspect all cases of alignment given that a proper display of alignment is rather unmarked. For our assessment activity, raters' attention can be directed to the display of alignment in response to initiations of our five main activities that we expect the test takers to complete for the task—SI, WT, CE, PA and AT—all of which we have discussed above. When recipient alignments are done properly, it would result in a streamlined management of those actions. In other words, the actions would be co-constructed competently and effectively by both or all parties involved.

Among the previous excerpts we have already discussed, example cases of successful alignment displays can be found in Excerpts 5.1 (SI), 5.9 (WT), 5.13 (CE), 5.18 (PA) and 5.23 (AT).

In the dataset, poor recipient alignment or recipient misalignment became salient when recipients showed insufficient uptake of activity recognition, which tended to result in the focal activity being derailed or even abandoned. For example, in Excerpt 5.29, a severe case of recipient misalignment occurred during an episode of CE talk between Thor, a game designer from Nintendo, and Panu, a real estate agent based in the United Kingdom.

Excerpt 5.29 *"No. No no…"*

```
407→ Pan:  is he:: interested in:: having a new headquarter?

408        if in that case you can +(1.0) do a business with me,=
     Pan:                          +RH point to himself

409        =because +(.8) +that is my goal for this um: welcoming party.=
     Pan:           +GZdown+GZ→Tho

410        =to find a new business partner.

411→ Tho:  no >no +no< it's not that.=
     Tho:         +BH waving

412→       =my: +(.6) my goal here is to find a new,
     Tho:       +BH↓↑, shake head

413        +(.8)
     Tho:  +shake BH↓↑

414  Tho:  soft- ↑UY game (.) com- company.

415        and we can make a- +some: collaboration with each others=
     Tho:                      +BH↺

416        =like <we can +rent> rent our company +character to +them,
     Tho:              +BH point to himself                  +BH put out in front
     Ath:                                        +RH reach out to Tho

417        and: >something like +that<
     Tho:                        +BH point to himself
```

Panu initiated his CE project to Thor (lines 407-409), asking if Thor knew whether his boss at Nintendo would be interested in building a new headquarters. Panu's initiation of CE was sharply rejected by Thor in line 411. Further examination reveals Thor's interpretation of the task as only permitting CE from his prepared goals for the role that he was playing. So, while Panu cited his goals (line 409) as the motivation for his proposed CE activity, Thor was orienting

more to how it was different ("*no nono… it's not that*") from his goals (lines 411 - 412). Therefore, instead of displaying an alignment to Panu's CE initiation, Thor's move to strike it down citing their conflicting goals as the reason for his action showed how his and Panu's actions were misaligned, resulting in an unsuccessful construction of CE.

Students' ability to display proper alignment in recipient actions can also be inspected in how they manage any disaligning actions. While *mis*alignment refers to problematic alignment between two adjacent actions and is something that we take as an evidence of interactional incompetency, *dis*alignments, on the other hand, are normal occurrences which sometimes happen when the recipient actions were designed to go against the agendas projected by the prior actions (Butler, Danby, & Emmison, 2011; Clayman, 2010).

PA activity is rife with opportunities for students to display alignment and disalignment. In this dataset, the majority of the students when presented with an invitation or suggestion to a post-conference activity opted for accepting or aligning with the party initiated the invitation. In a small number of cases, we can see some evidence of student IC through the way they managed to decline another group member's proposal. In Excerpt 5.30 below, during the talk in which participants shared with group members their plans for post-conference activities, Meiju displayed a competency to manage a declination to Matt's suggestion.

Excerpt 5.30 *"You should travel around"*

```
164→  Mat:   +how about you
      Mat:    +GZ→Mei

165           (.3)
166   Mei:    [I'm- ]
167   Mat:    [where] are you going.
168→  Mei:    .h I will::: (.) have to (.) go to the airport=because
169           I:: (.) have to (.) fly back to (.) China suddenly.
170   Mat:    o::h (.) it's nice city here.
171           (.4)
172→  Mat:    you should (.4) travel around.
173           (.6)
174→  Mei:    yes I'm very sad but (.2) my::: husband's mum is (.3) sick,
175           so I have to (.4) go back to see her.
176   Mat:    I'm sorry for that (x x).
```

At the beginning of Excerpt 5.30, Matt turned to Meiju, nominating her to take the next turn in line 164. The mini pause in line 165 prompted Matt to repair his question from "*what about you*" to a more specific question "*where are you going*", showing his orientation to a possible problem in Meiju's understanding of his original question. However, his repair came as an overlap to Meiju's turn initiation in line 166. She restarted her turn in line 168 with an audible in-breath and an elongated modal verb *will* which further delayed her upcoming response. With this, she packaged her response as a dispreferred one and went on to cite her previous engagement in China for her inability to go on travelling after the conference like other participants.

Despite Meiju's expressed plan to skip on any post-conference activity, Matt later issued a suggestion for her to travel around in line 172 showing another attempt on his part to pursue a positive response from her. She managed another disaligning response, using an agreement token to preface her dispreferred response to his suggestion, "*yes I'm very sad but ...*" in line 174. Meiju displayed an ability to manage her response to soften her refusal and maintain her affiliation with Matt. This refusal sequence added another layer of contingency to the PA activity and required students to showcase more kinds of interactional work compared to a preferred response sequence in accepting invitations (like the ones in excerpts 5.18-5.20). Given the perceived requirement of the task favoring the preferred response, it was rare in the data to see such dispreferred sequence, while in a real-life situation, students need to be competent in doing both.

So, in such case in which a response was disaligning, it is important in the context created for this task that students can also mitigate or minimize the potential disaffiliation implicated in their action. To discuss further the subject of recipient actions, the next section will introduce the

concept of affiliation, the nexus between the aspects of alignment and affiliation. The part of the discussion on alignment which overlapped with affiliation will continue below.

Affiliation (AF). Stivers et al. (2011) described affiliative responses as a "pro-social" action by ways that they are designed to match, support, and empathize with the preference of the prior turn. Within contexts that favor social solidarity as the outcome of the talk, actions which maximize affiliation are treated as preferred actions, while actions that can potentially promote destruction to social solidarity would be deemed dispreferred, and interactional resources would be mobilized to mitigate and minimize its disaffiliative impact (Lindström & Sorjonen, 2013).

Particularly in doing socializing activity such as our roleplay task where affiliation is considered the norm, the display and maintenance of affiliative stance would become categorically relevant for all participants at all times while performing for the task. Within all group performances, the students displayed a clear preference for building and maintaining social solidary as we can see the evidence of affiliative work permeated in almost every step of the way through a variety of verbal and non-verbal resources they had at their disposal.

Some considerations which raters can follow while judging the quality of students' affiliation display can be summarized into two main parts: first, its sequential placement in relation to the types of action in the first pair part, and second, the extent to which the turn design fits the sequential requirements demanded by the action in the first position. The students in this dataset tended to be able to identify sequential openings where affiliation display would be interactionally relevant. Their executions, however, varied quite greatly depending on their ability to pinpoint the target affective stance to which they should design their response to match and the range of interactional and linguistic resources they had available.

For instance, Excerpt 5.31 illustrates how Tim, Lee and Amber responding to Koe's statement about his plan to return to his hotel after the conference.

Excerpt 5.31 *"You might be tired"*

```
360→  Amb:   what about you guys?

361           +(.6)
      Amb:   +RH point to Koe

362   Koe:   .tsk (.) yeah I want to (.8) go back (.) to my hotel.
363          (.4)
364→  Tim:   [wow!
365   Koe:   [(x x)
366→  Amb:   hhhoh [ye:s.
367→  Lee:         [OH HHHH sound great!
368→  Amb:   okay (.) you might be tired
```

Because Koe's turn was formatted with a turn beginning "*tsk*" after a slight delay, it was hearable as a delivering of something with negative affective stance or at least was marked as not being newsworthy. Sequentially, by expressing a negative affective stance, Koe's turn provided something for other participants to align or affiliate with in the next turn. After a small delay, Tim was the first to provide affiliative response token "*wow!*" in line 364. Amber later supplied a change-of-state token "oh" pre-packaged in a laughter token (line 366) before issuing a receipt token "*ye:s.*" to further display her understanding of Koe's statement. Amber's receipt token came in an overlap with Lee's response action in line 367. Lee also provided a change-of-state marker "OH" and some laughter token in his response, but his action differed from Amber in its much louder volume and utilized a prosody which added a layer of excitement when he later produced a positive assessment "*sound great!*" at the end of his turn. Amber later added another layer to her recipient action in line 368 with "*okay*", a display of acceptance and understanding, followed by a comment "you might be tired" offering a candidate understanding of why he wanted to return to his hotel instead of going sightseeing like others.

One stipulation should be made about a possibly ambiguous interpretation the analysis can make about the students' laughter in lines 366, 367 before proceeding with the analysis. Since Koe's statement was produced as part of a series of talk about future plans after the

conference, his plan to simply go back to his hotel could be seen as an unexpected decision which could invite some laughter from his peers. For this reason, it was unclear if their laughter should be considered as part of their on-task or off-task interaction.

However, if we only consider the context within the roleplay as the only available context, it would become quite apparent that Tim and Lee's displays of their affiliation to Koe's negative news delivery were not really a good match to the affiliative stance in Koe's turn. The only response action which display appropriate affiliative stance to Koe's was Amber's "*okay… you might be tired*" in line 368.

Many researchers have suggested that the management of affiliation and alignment are often interrelated (Butler et al., 2011; Steensig, 2012), and therefore, in some cases, our evaluative decision regarding the quality of affiliation display must be considered in relation to their display of alignment and vice versa.

Within the types of actions relevant in our roleplay, the relationship between the alignment and affiliation in recipient actions hinges on the concept of preference organization (Pomerantz, 1984; Schegloff, 2007), which describes sequence organizations of preferred and dispreferred responses. For the type of actions which make relevant more than one possible alternative responses, not all possible responses are equally valued (Schegloff & Sacks, 1973). Depending on the kinds of action in the first pair part and the linguistic configuration in which that action is formatted, different types of response would embody different alignment of the recipient toward the prior's turn action projection. While the terms alignment and preference are different, the former being used at the level of action or sequence type and the latter with finer grained account which considers anticipated response controlled by different turn designs, the nuance differences between the two terms based in the study of L1 language use are not directly relevant here. So, to simplify the current study's recommendations for rubric design and raters'

training, preferred responses are used to display alignment and correspondingly, dispreferred responses are used to display disalignment. Normative organization of action sequences of both preferred and dispreferred responses is designed to promote social solidarity and therefore constitute a display of affiliation in recipient actions.

Schegloff (2007) provides a very useful list of regularly produced features which have come to characterize a response as either preferred or dispreferred. Preferred responses tend to get delivered in a "normal" transition space (no mitigation) and are likely to be less elaborated, short, and to the point. In contrast, dispreferred responses tend to be attenuated, mitigated with delays, and commonly delivered with more elaboration such as explanations, excuses, hedges, or disclaimers, i.e., "*I don't know*".

Raters can refer to this list of recipient design features in evaluating students' display of affiliation. In employing the appropriate preference-type response, students can be seen to display a proper recognition of action in the first pair part. At the same time, employing aforementioned features in their preferred or dispreferred responses can be taken as an ability to display recipient affiliations.

For example, during an episode of Ann's WT in Excerpt 5.32, we can apply what we adopted from Schegloff (2007) into noticing some of Ann's ability to manage recipient affiliation.

Excerpt 5.32 *"um… I don't go"*

```
203    Ale:   so it's mean you can (.3) build a s: (.) aerospace?
204           (.3)
205→   Ann:   ↑ye:s.
206    Hug:   wo::::[:::w ]that's amazing,
207    Kai:         [wo:w.]
208    Chr:   can you take us (.) there?
209    Kai:   [h h [h
210    Ale:   [ah h[hh
211    Chr:        [to Mars or something?
212           (.2)
213→   Ann:   °um:::° >I don't go I have< +(.3)
       Ann:                                +BH put up in front

214           I just (x) go +I- I can't-
215    Chr:   ah H
```

```
216    Hug:   ah::
217    Chr:   I see::,
```

Here, after Alex formulated an upshot of his understanding (line 203) up for Ann's confirmation, Ann's response in line 205 was clearly formatted as a preferred response evident in the precision of her answer and the minimal pause before the response was supplied. In Ann's other response later in line 213, she started her turn with an elongated "*um:::*" produced in a quiet voice as her way of delaying her dispreferred response before stating what sounded like an account to explain why she is incapable of complying with Christian's prior request. From this available evidence, Ann had displayed an awareness of an organization of preferred and dispreferred sequences in ways which conform part of our standard for affiliation and alignment display.

Conclusion

This chapter presents the findings from CA used to describe interactional actions and activities nominated to represent the students' interactional competence construct for their performance on the *socializing task*. The productive activities which students consistently performed in the roleplay data include a self-introduction, a work-related topic discussion, a contact exchange, a post-conference planning, and finally the closing of the roleplay activity. For the recipient actions, the study identified three aspects of response management including understanding display, alignment display, and affiliation display. In the next chapter, the results from this qualitative analysis will be discussed to address how these findings can help us answer the qualitative research questions set out in Chapter 4.

CHAPTER 6

DISCUSSION OF QUALITATIVE RESULTS

This chapter discusses the research findings in relation to the qualitative Research Questions 1-3 which the study proposed at the beginning of the study. In order to examine the validity evidence (Messick, 1989b) of the proposed assessment instrument and the extent that it could generate meaningful scores from observations of test takers' interactional competence (IC), the discussion begins with the identification of interactional phenomena from task performance data and the how the current study operationalized IC into a measurable assessment construct. Then, this study will address the construction of a rating scale which constitutes a critical link between the scores and the observations of test takers' IC in performing this *socializing task*.

Research Question 1: *From the students' roleplay performance data, what are the constitutive interactional phenomena in the form of actions or courses of action which can be established as the targets for comparing IC across the dataset?*

To establish the basis for cross-sectional comparisons needed for measuring purposes, the microanalytical lens from Conversational Analysis (CA) were adopted in identifying these actions and courses of action. The qualitative chapter provided an analysis of the five key activities—self-introduction, work talk, contact exchange, post-conference arrangement, and activity termination—that were narrowed down as the focus of our study from the entire roleplay performance data on the *socializing task*. Additionally, three recipient actions—understanding display, alignment display, and affiliation display—were included to provide a complementary account to the dominantly productive skills observed in the five main activities. These eight

actions altogether formed the core activities on which raters would assign interaction competence (IC) scores representing the student test takers' IC construct in performing the *socializing task*.

The five production actions focus on the students' interactional competencies in managing action initiations of the selected five interactional activities. The qualitative analysis described typical sequence organizations of each activity and the interactional components each activity typically encompasses. Based on these qualitative CA findings in Chapter five, the actions and course of actions which have been included to represent the target construct in each production activity are summarized in Table 6.1 below.

Table 6.1
A Summary of CA Findings of Typical Sequence Organization for the Production Activities

Production Activity	**Typical course of actions**
Self-introduction (SI)	- Initiating SI sequence through topic transition - Managing their SI sequence with an audience - Closing their SI sequence by selecting the next speaker or nominating the next topic
Work talk (WT)	- Initiating WT sequence through topic transition or topic proffer - Sustaining WT talk through initiating post-expansion sequences
Contact exchange (CE)	- Invoking CE activity by transitioning from WT when appropriate - Initiating a request, an offer, or a suggestion sequence with an audience to establish a business connection
Post-conference arrangement (PA)	*PA as a spontaneous joint activity* - Issuing a suggestion or request sequence for a joint activity (designed to appeal to the audience) after two or more students established that their plans overlapped - making arrangement for the joint activity, i.e., meeting time or activity details or, *PA as a prearranged occasion* - Issuing an invitation sequence (designed to appeal to the audience) - making arrangement for the joint activity, i.e., meeting time or activity details
Activity termination (AT)	- Initiating a pre-closing sequence - Coordinating the conversation closing sequence - (optional) Engaging with co-participants in a minimal last business sequence - Terminal exchanges

In the part of recipient actions that the current study has included in the operationalization of the IC measurement construct, the qualitative analysis has provided descriptions of the sequential shapes and environments in which students in the performance data displayed evidence of competencies in the three recipient actions —displaying understanding, displaying

alignment, and displaying affiliation. While our descriptions of the productive activities are specific to each activity, this study's treatment of the three recipient actions included in the measurement construct is not. For each action under recipient actions, students would be evaluated based on their ability to competently display understanding, alignment, and affiliation overall throughout their roleplay performances.

In the process of narrowing down the analytical focus in order to identify comparable interactional phenomena at this stage, the study was faced with some challenges in strictly applying the CA's methodological apparatus when applying such procedures with the goal to do cross-sectional comparisons with assessment purposes in mind. The longitudinal comparative studies (i.e., Hellermann, 2011; Pekarek Doehler & Berger, 2016) were able to warrant their claims of IC development through tracing the trajectory of changes within the established collections of specific actions which requires similar interactional work (e.g., a self-initiated story opening told in the first position in Pekarek Doehler & Berger, 2016). Schegloff (1993) recommended that such precision is necessary for any analysis which seeks to quantify interactional phenomena, either through counting, coding, or putting them up side by side for comparative purposes.

However, adopting such level of precision can be impractical for the kinds of interactional data and the measurement agenda that this study is pursuing. First of all, the performance data were elicited by a very broadly defined task permitting students to choose or design their own roles as well as the imagined relationships between these roleplay characters. With an unstable set of participants, establishing a consistent collection of actions which requires similar interactional work can be very difficult to attain. Secondly, to assign scores or pass judgments on students' IC in these precise and specific actions needs a fine-grained analysis of

student interactional performances which requires a great amount of time and raters' training to facilitate such practice.

To balance the concerns over precision on one hand and practicality on the other, the current study has chosen to identify for each productive activity a cluster of key actions which roughly form an activity which can be recognizable by raters whether they have a background in CA training or not. While the practice adopted in this study is still at an exploratory stage that may still be far from being optimized due to the practical demands set by the number of test takers, available resources, and the length of each performance on the *socializing task*, for a valid and defensible observations to represent the test takers' IC in performing that specific action or course of actions, the current study argues that following Schegloff's (1993) recommendations is necessary for any future assessment instrument developments targeting IC as its construct.

Research Question 2: *What are students' methods in varying degrees of success, in accomplishing the actions or courses of action identified as the targeted interactional phenomena?*

The qualitative findings of "member methods" via conversation analysis (CA) of student performances on the five productive activities and three recipient actions have been compared to the existing body of knowledge that CA has to offer on "normal" behaviors on closely approximate actions. While findings from CA literature can provide a baseline for a set of standards and expectations the rubric can later adopt as its assessment criteria, a detailed investigation into the students' methods of participation can help reveal the interactional problems and shortcomings which can be used in scaling our observations from high to low levels of interactional competencies specific for each action and activity and more importantly specific to the kinds and abilities of students performing this assessment task.

Table 6.2

Descriptions of Successful and Problematic Managements of Each Activity

Activity	Descriptions (Successful)	Descriptions (Problematic)
Self-introduction (SI)	- natural topic transitions - display recipient design in managing their SI - display membership knowledge of the role they play - natural topic closing and nominating the next speaker in a timely manner	- lacking transition - unorganized sequence of actions - poor turn designs - lacking resources to initiate repair sequences when a repair initiation is necessary
Work talk (WT)	- natural topic transitions or natural initiation of a new topic - active participation on developing WT talk co-participants - good organization of WT sequence - coherent turn composition which fit well with context generated in prior talk - invoke professional narrative in their WT	- little to no topic transitional devices - non-conventional methods in recipient selection - orientation to pass turns quickly rather than expanding the topic - lacking active participation in developing WT with co-participants
Contact exchange (CE)	- natural topic transitions - recognize appropriate interactional opportunities to initiate CE sequence - able to complete CE or pursue the activity even when the CE sequence get derailed	- lacking active participation in developing CE with co-participants - misplace CE within the overall sequential position of the roleplay performances - off-target composition of CE - fail to complete CE or pursue the activity when the CE sequence get derailed
Post-conference arrangement (PA)	- recognize appropriate interactional opportunities to initiate PA sequence - display recipient design in managing their PA to maximize the change for desirable outcomes - display awareness of the sequential organization of PA activity - able to design their actions to fit the ongoing progressivity of the group's talk	- lacking active participation in developing PA with co-participants - lacking awareness of the sequential organization of PA activity - poor design of their PA activity resulting in a poor fit to the ongoing progressivity of the group's talk
Activity termination (AT)	- recognize appropriate interactional opportunities to initiate AT sequence - natural management of pre-closing and closing sequence organizations with co-participants - formulating "reason for leaving" clearly and effectively - reinvoke appropriately a minimal last business sequence	- initiate AT sequence disruptively - lacking awareness of the sequential organization of PA activity - poor pre-closing turn design and execution
Display understanding (U)	- appropriately display and maintain intersubjectivity with co-participants throughout the roleplay performance	- failure to address or resolve conversational troubles when a repair is due
Display alignment (AL)	- display recognition of the action or activity is initiated in a timely manner - able to contribute meaningfully to the action or activity in progress - display flexibility to handle both aligning response and disaligning response in sequentially appropriate manner	- slow to show any uptake in recognizing the action or activity being initiated - display misalignment (incorrect recognition of the action) - unable to
Display affiliation (AF)	- able to diversify interactional resources in displaying affiliative stance - able to match or meet the expected level of affiliative work demanded by the situation	- limited resources in displaying affiliative stance - unable to match or meet the expected level of affiliative work demanded by the situation

Based on the analysis of successful versus problematic cases of students' management of each activity, Table 6.2 above provided a summary of the study's observations of students' high level of IC in each activity and vice versa.

What we can identify is a pattern where successful management of these eight activities requires two major components: a displayed awareness and control of the sequence organizations in each of these activities and an ability to design and compose turns and sequences cohesively to what came before and effective in carrying out the actions.

A challenge at this stage of rubric construction is how to ascribe different values to the less successful performances on these eight activities. While overall findings on IC development pointed to a more effective accomplishment of interactive practices over time, Nguyen (2012a) noted that IC development might not be linear, and the developmental paths can differ from one person to another (p. 229). To address this challenge, we must also take the raters' behaviors into account as they are the people who would be using the rubric to assign scores to each of the student performances. Relevant questions also include how many levels of accomplishment raters can distinguish reliably and what each level of accomplishment constitutes. A provisional rubric was created to explore these fundamental questions which are necessary for advancing assessment movement of IC.

Research Question 3: *How can the rich description of students' task performance inform the data-driven construction of the IC assessment rubric?*

A rich description of how students performed the task fosters more precise and better-fitted criteria descriptors, the conditions which a data-driven approach to rubric construction argued would potentially contribute to better reliability of a test instrument and usefulness in providing test users with directly relevant feedback (Fulcher et al., 2011). Given that our goal is

to devise a rubric or a rating scale which can be used by the general language teaching professionals, who were the original raters of the *socializing task*, there is a need to simplify the descriptions to maximize a uniform interpretation of the rubric descriptors when it is being used.

Advantages of the construct of IC are that it is a competence to execute social actions locally in real time (Hall et al., 2011; Mehan, 1979) and that IC cannot be separated from its performance (Roever & Kasper, 2018), so there is little need to make inference about students' IC outside of what is observable in their performances. It has been decided, therefore, that the rubric descriptors contain, as much as possible, descriptions of observable behaviors or accomplishments for each selected interactional activity.

The scoring system adopted at this stage of the study is on a scale of one to five: "1" being the most problematic case of executing the actions and "5" being the target-like performance. Based on the observations of these problematic cases, more disruptive operations of these activities were those cases in which students failed to organize their activity in a sequentially appropriate way. The middle category "3" represents this step when the executions are identified as having a poor sequential organization. So, if the students have managed their actions with appropriate sequence organization, they should at least get a "4" or a "5" rating, depending on the design and compositions of their turns and actions. The "0" category has also been included in a case that students did not make any attempt to initiate any of the productive activities at all.

A summary of recommendations emerged from the above analysis for assessing each of the eight activities can be found in Table 6.3 below.

Table 6.3

The Proposed Rubric for Assessing IC of the Targeted Task

Item	Actions	Sequence organization:	Action composition:
1	Self-introduction (SI)	• transition appropriately from previous action into the self-introduction • complete their own SI project • actively take part in progressing the roleplay by nominating the next speaker (including self-nomination) or initiating sequence closure before transition to the next activity • initiate repair when potential understanding problems occur	• having prepared sufficient resources about their role to participate in a "complete" self-introduction sequence • display understanding of target domain (a business conference) through the design and execution of their role • display ability to perform SI in a routinized, quick paced sequence
2	Work talk (WT)	• appropriately initiate WT sequence or transition from previous topic into the work/job responsibility talk • actively work to develop the topic of work talk (own WT or peers' WT) (not transition away abruptly) • conclude their own WT by nominating the next speaker or transitioning into a different topic when appropriate • initiate repair when potential understanding problems occur	• having prepared sufficient resources to talk about their job, company and work responsibilities • display sufficient understanding of the job position and field of industry chosen for their role through the design and execution of their WT • construct a professional narrative of their company, their job position and their purpose they were there at the conference
3	Contact exchange (CE)	• appropriately pivot ongoing talk into creating opportunity for CE • invoke 'reasons for CE' before launching any speech acts which chosen for the activity: i.e., request / offer / suggestion. • display the ability to re-invoke the CE even when the sequence progressivity got sidetracked. • initiate repair when potential understanding problems occur	• clear audience selection and cater the design to maximize positive response • in the request / offer / making suggestion sequence, cite appropriate reasons for their respective companies to work together • display sufficient understanding of the job position they hold in their designs of 'reasons for CE'
4	Post-conference arrangement (PA)	• appropriately initiate PA activity or transition from ongoing talk (non-work talk) into PA activity • display awareness of sequence organization of invitation and able to initiate it appropriately • display ability to provide account for decision making (accept invitation or decline invitation) in sequentially appropriate way • initiate repair when potential understanding problems occur	• the invitation includes sufficient descriptions of the proposed future activity or place they want to visit • the proposed descriptions were designed to appeal to the audience • able to design their turns with rooms for plan adjustment • build up the decisions of PA as the activity progresses without abandoning prior talk when making new plans
5	Activity termination (AT)	• initiate AT after previous topic has already been closed (not disruptively invoke AT while other topic is still in development) • provide reasons for leaving before excusing oneself out of conversation • display awareness of sequence organization of AT (reason for leaving >'nice to meet you' tokens > possible last business > goodbye tokens) • initiate repair when potential understanding problems occur	• clearly orient to activity termination through pitch change, body disposition, turn design, etc. • display ability to perform AT in a routinized, quick paced sequence
6	Display understanding (U)	• provide acknowledgement tokens where it's sequentially needed • address all possible misunderstandings through employing appropriate types of repair	• use appropriate acknowledgement tokens (head nod or other embodied acknowledgement tokens are okay)
7	Display alignment (AL)	• provide preferred response (agreement or disagreement tokens) where it's sequentially called for	• use agreement/disagreement tokens appropriate for the context
8	Display affiliation (AF)	• provide positive assessment tokens where it's sequentially needed	• use affiliation tokens appropriate for the context • possess a variety of lexical resources for affiliation displays

CHAPTER 7

QUANTITATIVE RESULTS

This chapter presents quantitative findings from a FACETS analysis based on the use of the proposed rubric in assessing student interactional competence in performing the *socializing task*. This chapter is divided into three main parts. First, before we discuss the main findings, preliminary results including descriptive statistics, correlation analyses, and a check for test unidimensionality are presented. The second section reports the measurement results from FACETS, discussing the overall model and each of the facet's results in separate measurement reports. In the third and final section, the results of FACETS interaction analysis between each facet are presented to explore possible biases from different raters and items included in the measurement model.

Descriptive Statistics and Correlation Analyses

From the remaining set of 148 students after the data cleaning included in the quantitative analysis, their raw scores on the eight items representing five production activities (Self Introduction, Work Talk, Contact Exchange, Post-conference Arrangement, and Activity Termination), and three recipient actions (Understanding, Alignment, and Affiliation) assigned by the six raters were summarized in Table 7.1 below.

Table 7.1
Descriptive Statistics

Items	Rater 1		Rater 2		Rater 3		Rater 4		Rater 5		Rater 6	
(*N*=148)	***M***	*SD*	***M***	*SD*	***M***	*SD*	***M***	*SD*	***M***	*SD*	***M***	*SD*
Productive items												
SI	**3.49**	0.88	**4.42**	0.78	**3.11**	0.69	**2.97**	0.90	**3.57**	0.93	**3.78**	1.05
WT	**3.45**	0.86	**4.09**	0.95	**3.09**	0.81	**3.32**	1.41	**3.72**	0.92	**3.67**	1.18
CE	**1.45**	1.48	**1.91**	2.04	**1.76**	1.60	**1.43**	1.51	**1.95**	1.63	**1.82**	1.85
PA	**3.25**	0.75	**3.82**	1.69	**3.01**	0.89	**2.97**	1.36	**3.36**	1.24	**2.87**	1.60
AT	**2.80**	0.93	**3.28**	1.48	**2.66**	0.89	**2.25**	1.15	**2.95**	1.37	**3.20**	1.36
Recipient display items												
U	**3.52**	0.71	**4.80**	0.49	**4.81**	0.39	**3.15**	1.33	**4.66**	0.64	**3.78**	1.53
AL	**3.35**	0.71	**3.45**	1.66	**4.48**	0.67	**3.13**	1.32	**4.02**	0.89	**4.09**	1.08
AF	**2.78**	1.01	**2.98**	2.02	**3.61**	0.69	**3.16**	1.32	**4.48**	0.73	**3.27**	1.74

From inspecting the means and standard deviations for each item, the item Contact Exchange (CE) received the lowest mean scores consistently from all 6 raters ranging from 1.45 (Rater 1) to 1.95 (Rater 5), suggesting that CE is the most difficult among all the actions students were required to complete. In terms of score spread, CE also has the highest degrees of dispersion across all raters. The high spread of scores on CE signifies that even though the students overall were struggling with this activity, they were not all consistently bad at doing CE. Across this group of students, some managed the activity successfully, and some failed to do so.

Apart from CE, the pattern of the difficulty of other activities seems unclear. Four out of six raters (all except raters 4 and 6) have given the highest scores for Understanding (U), ranging from 3.52 (Rater 1), 4.80 (Rater 2), 4.81 (Rater 3), and 4.66 (Rater 5), making item U potentially the easiest item out of all the rating criteria. Among the raters whose mean scores for U were their highest, all their standard deviations are low, between 0.39 to 0.71, meaning that they were all quite lenient in assigning higher scores across all participants for this item. In contrast, for raters 4 and 6, who appeared much stricter in grading item U, their assigned scores were more widely dispersed (1.33 and 1.53, respectively). This could mean that they have more gradations in distinguishing between higher and lower degrees of achievement in the student performances on item U more than the other four raters.

In order to gather preliminary observations on the relationships between these eight items, bivariate Pearson correlation coefficients were calculated and reported in Table 7.2.

Table 7.2
Inter-item Correlation Matrix Based on Composite Scores

Pearson Correlation (*r*) (*N*=148)	SI	WT	CE	PA	AT	U	AL	AF
Productive items								
SI	1	.403**	.331**	.227**	.293**	.449**	.458**	.524**
WT		1	.103	.380**	.378**	.482**	.509**	.502**
CE			1	.243**	.408**	.372**	.284**	.337**
PA				1	.352**	.343**	.468**	.449**
AT					1	.491**	.502**	.508**
Recipient display items								
U						1	.814**	.810**
AL							1	.815**
AF								1

** Correlation is significant at the 0.01 level (2-tailed).

All items displayed to correlate in the same direction given that all the correlation coefficients reported here are positive. Furthermore, almost all of the correlations are significant given the size of the dataset (*N*=148), except the correlation between WT and CE that is not (i.e., *r*=.10). When we look into the correlations between the remaining productive items, even though their correlations are significant, the reported Pearson correlation coefficients are rather weak, with *r*=.41 or lower. The correlations between the items which targeted recipient actions, on the other hand, are fairly strong at *r*=.80 or above. The relationships between productive items and recipient display items are moderate, except the correlations with CE that are notably low. All in all, the recipient display items are fairly more uniform compared to the productive items that appear to differ in the construct they target. More findings from FACETS analysis should help confirm or further explain these observations.

We also explored the degree of raters' agreement through inspecting the relationships between the scores they assigned across all of the eight items. The bivariate Pearson correlation coefficients which also serve as the pair-wise interrater reliability estimates have also been calculated and reported (see Table 7.3).

Table 7.3
Interrater Correlation Matrix

Pearson Correlation (*r*) (*N*=1148)	Rater1	Rater2	Rater3	Rater4	Rater5	Rater6
Rater1	1	.551**	.540**	.498**	.487**	.445**
Rater2		1	.461**	.423**	.452**	.389**
Rater3			1	.468**	.653**	.458**
Rater4				1	.455**	**.347****
Rater5					1	.429**
Rater6						1

** Correlation is significant at the 0.01 level (2-tailed).

These interrater reliability estimates are not very high, which strongly suggest that the current version of the rubric should be rated by more than one rater to warrant the scores' reliability. However, these correlation coefficients represent only the reliabilities of a single set of ratings. To calculate the current study's interrater reliability estimate adjusting for the fact that there are actually six raters, the Spearman-Brown Prophecy formula (i.e., $r_{xx'}=n(r)/((n-1)r+1)$

when *n*=number of raters and *r*=correlation between two raters) can be used to provide the interrater reliability estimate (Brown, J. D., 2005). Based on correlation coefficients shown in Table 7.3, the lowest correlation is found between Rater 4 and Rater 6 (*r*=.347). Using the Spearman-Brown Prophecy formula, an interrater reliability estimate for the rating practice in this study is 0.76 (i.e., (0.347x6)/(5x 0.347)+1)), or 76 percent. Because we have based our calculation on the lowest correlation between our raters, this reliability estimate generated through this calculation provides a conservative estimation of the reliability, meaning that the reliability of 0.76 is unlikely an overestimation of the actual interrater reliability of this study's rating procedure overall.

To conclude the findings based on the descriptive statistics and correlation analyses, raters appear to differ in their severity when it comes to student performances in different items. Unfortunately, the information about who these students were and the characteristics involving how well they managed other items were not available through interpreting these results. The same applies to the raters' behavior and how they interact differently when assigning scores for different items. However, these questions can be taken up when we turn to the FACETS analysis discussed further below.

Checking for Unidimensionality

Before proceeding to the discussions on many-facet Rasch analysis from FACETS, the dataset was checked for whether it met the requirements for using the Rasch model. Fundamental to all models under Item Response Theory (IRT), we should not attempt to measure more than one theorized construct at a time. For the Rasch model to reveal meaningful and interpretable information, the measurement we wish to study must be unidimensional as it would otherwise be unclear which latent construct the Rasch model actually operationalizes (Linacre, 1998a).

Fulfillment of unidimensionality condition is a matter of degree (Bejar, 1983), and factor analysis can be used to help detect the possibility of multiple dimensions within the data. Based on the composite scores derived from the average scores of each item from all six raters of the remaining 148 student participants, Principle Component Analysis (PCA) was performed and the results showing eigenvalues of the PCA extraction are presented in Table 7.4, and a scree plot which provides a visual representation of the relationship between eigenvalues and the number of components is presented in Figure 7.1.

Based on the Kaiser stopping rule, which indicates that only components with an eigenvalue greater than one should be considered in the analysis (Brown, J. D., 2009), the scores produced by eight items were shown to have only one component, with the eigenvalue greater than 1.00. Another approach for determining the number of factors is the scree test, which considers the magnitude of eigenvalues and suggests the cut point for the number of extracted component being a point where the eigenvalues precipitously drop and level out. For this dataset, both the scree test and Kaiser's stopping rule provide the similar conclusion that there is one factor, and this factor accounts for 52.38 percent of the total variance in the dataset, which warrants our assumption of unidimensional data needed for FACETS analysis.

Table 7.4
PCA for Composite Scores for eight Items on the Proposed Rubric

Total Variance Explained

Component	Eigenvalues		
	Total	% of Variance	Cumulative %
1	4.190	52.377	52.377
2	0.959	11.984	64.361
3	0.805	10.065	74.426
4	0.679	8.486	82.912
5	0.598	7.470	90.382
6	0.428	5.346	95.728
7	0.180	2.249	97.978
8	0.162	2.022	100.000

Note. Extraction Method: Principal Component Analysis

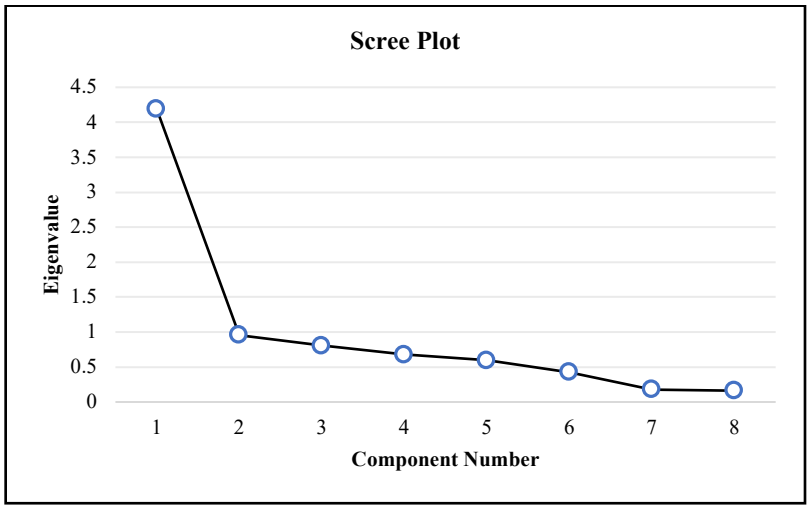


Figure 7.1 *Scree Plot for Composite Scores for Eight Items on the Proposed Rubric*

As we have seen earlier in Table 7.4, the first component accounts for 52 percent of the total variance in the composite scores. While the model is clearly unidimensional, we can further inspect the component loadings to see how much variance is accounted for by the Rasch model (see Table 7.5).

Table 7.5
Component Loadings from PCA Based on Composite Scores

	Loadings	Communalities (h^2)
SI	0.632	0.399
WT	0.657	0.432
CE	0.491	0.241
PA	0.583	0.340
AT	0.677	0.458
U	0.864	0.746
AL	0.880	0.774
AF	0.893	0.797
Cumulative Variance		4.187

Because FACETS assumes the dataset measures only one construct at a time, factor loadings of the first component can reveal in detail how much variance from each item has been accounted for in our study's measurement model. According to the first component's loadings displayed in Table 7.5, over 70 percent of the variance within items U, AL, and AF, all of which targeted at measuring recipient actions, were accounted by the model. On the other hand, the

model accounts for a much lower proportion of the variance in each of the productive actions especially items CE and SI, i.e., 24 percent and 40 percent, respectively. This pattern is to be expected given the correlation coefficients, shown earlier in Table 7.2, among the recipient action items that are much higher than the productive actions. This means that the results from the multifaceted Rasch could be more heavily accounted for by the student's recipient displays and less so by their productive actions.

FACETS Measurement Reports

Many-facet Rasch analysis was conducted through the computer program FACETS version 3.81.0 (Linacre, 2018). In specifying the parameters for FACETS analysis, each facet represents a variable which can influence the object of our analysis, that is, our test scores. In this study, three facets have been included. In the following sections, we first discuss the rater severity facet before moving on to the student ability facet and the item difficulty facet in turn.

One key advantage of FACETS is that it offers a graphic presentation of all facets under one 'ruler.' This provides a map of students' ability, raters' severity, and item difficulty all on the same scale. FACETS calibrates the effect of both raters' severity and item difficulty on students' scores and reports all the measures in *logit* units. The logit scores represent the student ability in terms of their likelihood to successfully complete the task. This means that when the student's ability matches the difficulty level of the task, he or she would have a 50 percent chance to either successfully complete the task or fail to do so. Because FACETS reports all measurements in this one common unit, analysts are able to directly compare student's ability, item difficulty, and rater severity all at once. Based on the scores produced by our six raters, Figure 7.2 displays the 'vertical ruler,' or a variable map, which summarizes all the facets in a single visual representation using the logit scale.

Conventionally, item difficulty and rater severity facets were set to center at zero logit to allow for the student ability logits to move freely. In this vertical ruler, the first column shows the measurement scale in the logit units used by all the facets in this study. The next three columns display the rater facet measurement, student facet measurement, and item facet measurement consecutively. Finally, as this study adopts the Partial Credit Model in analyzing the category measurements in each of our items, Columns 5-12 display the thresholds in climbing up the score levels for each item. Higher logit scores signify greater ability in the students, more severe judges, and more difficult items or tasks, while negative logit scores indicate lower ability students, more lenient judges, and easier items or tasks.

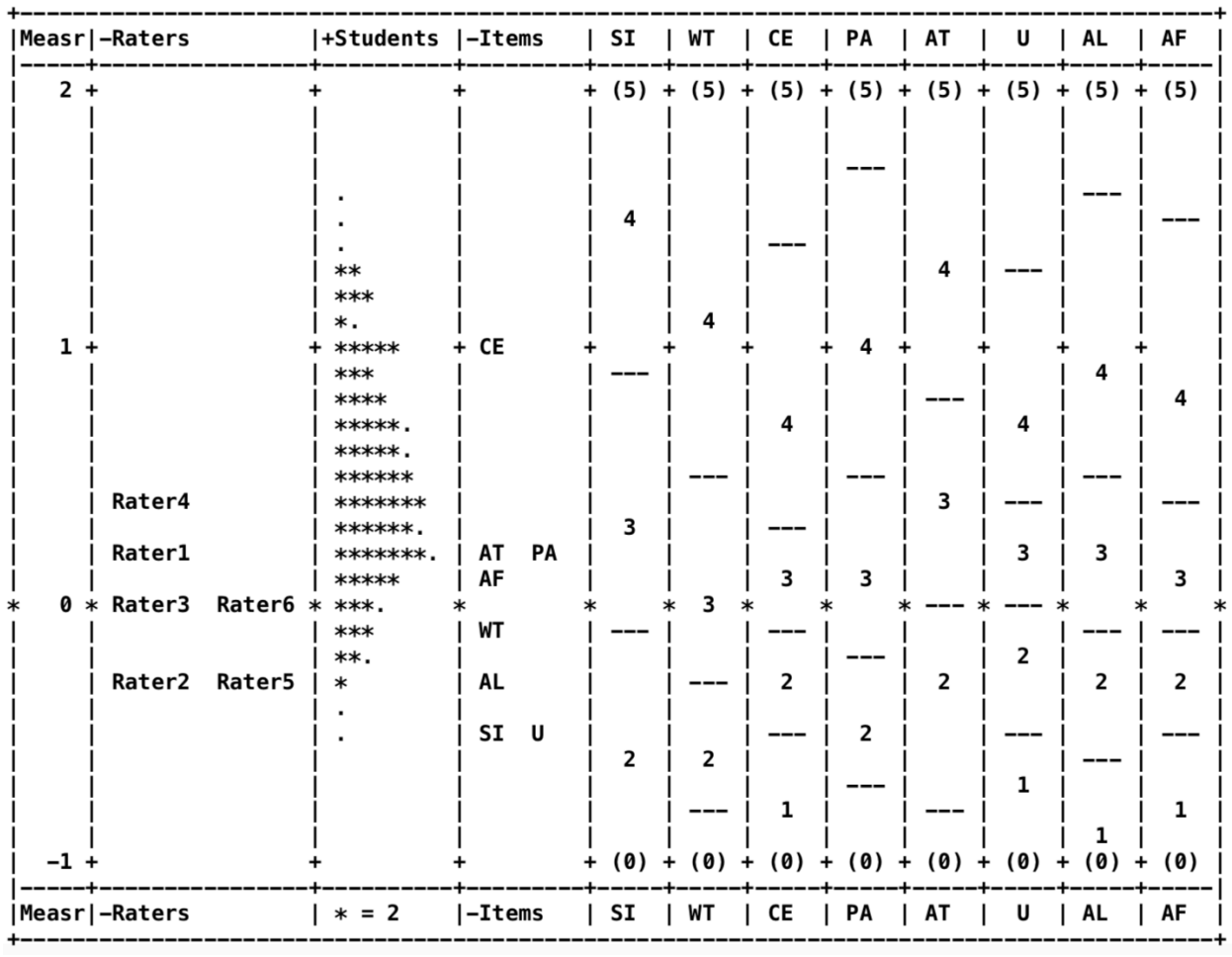


Figure 7.2 *FACETS Summary: All Facets Vertical Ruler*

In the second column, which shows the rater severity logits of all the raters, we can see that Rater 4 was the most severe, and Rater 2 and Rater 5 share the rank in being the most lenient raters. In the middle, we have Rater 3 and Rater 6 whose severities fell at about the same level at zero logit. In the third column of the vertical ruler, we can observe the spread of the student ability ranging from -0.5 logits to 1.5 logits. Each asterisk on the scale represents two students, and each dot represents one student. We can also see the placement of items along the vertical ruler in the fourth column. Among all the activities that we have included in our study, Contact Exchange (CE) appears to be the most difficult task as it ranks close to a full logit away from other items on the vertical ruler. Two activities, Self-Introduction (SI) and Understanding (U), shared the rank of being the easiest of all the eight activities given their positions below other categories on the vertical ruler. This information enhances the findings from descriptive statistics as it not only confirmed that CE is the most difficult activity for this group of students, but also that the item is more difficult than other items by close to one logit.

To maximize the test reliability, we generally want to see a match between the pattern of the student test takers and the pattern of item difficulty. According to what we can see from this overall facet map, it appears that there are some gaps in the student ability logit scale for the upper two thirds of the students' ability logits. CE was the only item that could differentiate students' ability in this upper range. If the goal of were to further distinguish students in this upper level, more items targeting students between 0.5 to 1.0 logits and ability greater than 1.0 logit should be added in the future versions of this test. Two items, AT and PA, may be redundant given their similar difficulty logits, but because of their placement at around the logits in which many students were also placed, having two items at this level of difficulty can improve the test reliability overall. The actual concern was the redundancy found between items SI and U, which were mostly too easy for the whole group of the students. Given that there were very few

students whose ability was targeted by SI and U, future versions of the rubric may discard one of these items to generate a truncated version of this test.

Column 5 show how the scale of 0-5 was utilized in each of the items from self-introduction (SI) in the fifth column, to work talk (WT), contact exchange (CE), post-conference arrangement (PA), activity termination (AT), understanding display (U), alignment display (AL), and affiliation display (AF) in the last column consecutively. The Partial Credit Model (Masters, 1982) utilized in this study independently estimated category measurement reports from 0-5 scales for each item, allowing for separate examinations into how the scales in each of the items may function differently. Based on the vertical ruler for the scales in each of these items, it can be predicted that the students at zero logits ability level are likely to get the score of three, which is the middle category, across all eight items. Different sizes of the categories on the vertical ruler indicate that each category might not function as expected, revealing that some of the scores might have been underused. This issue of rating scale category functioning will be addressed when we discuss the results of category measurement report below.

Rater measurement report. The more detailed results into how the raters behaved while using the rating scale we are currently studying can be found in FACETS' rater measurement report. Table 7.6 presents rater severity logit measurements, errors, and fit statistics for the six raters. We have earlier discussed the interpretations of logit measurements that for rater severity, the higher the logits, the stricter or tougher they are, and vice versa. In this study, the six raters' severity spread out between -0.26 logits (the most lenient rater) to 0.21 logits (the most severe rater). The high separation index of 9.45 further supports our interpretation that the raters were different in their severity. This separation index also provides an estimate that there are about nine statistically distinct levels of rater severities in this dataset. The significant fixed chi-square statistic (χ^2=461.8, df=5, p<.00) correspondingly rejected the null hypothesis that all raters

behaved similarly. For the reliability estimation, FACETS also reported the separation reliability of the raters in this study at 0.99, which is very high. By putting together all these indicators, this means that raters were reliable in maintaining their severity levels across the dataset.

Table 7.6
Measurement Report for Raters

Raters #	Severity Logits	Model *S.E.*	Infit MnSq	Outfit MnSq
Rater1	0.21	0.02	0.50	0.54
Rater2	-0.26	0.03	1.83	1.78
Rater3	-0.03	0.03	0.63	0.67
Rater4	0.36	0.02	1.05	1.11
Rater5	-0.25	0.03	0.89	0.88
Rater6	-0.02	0.03	1.33	1.29
M	0.00	0.03	1.04	1.04
SD	0.25	0.00	0.49	0.45

Note. Reliability = 0.99; Separation: 9.45; Fixed chi-square: 461.8 (df=5; p<.00), RMSEA = 0.03

FACETS also reported fit statistics which are very helpful in pointing out aberrant performances in each of the raters that should be further inspected in more detail (Bond & Fox, 2015). Two kinds of fit indices are reported in FACETS: the infit statistics and outfit statistics in the form of mean square indices. Both infit and outfit indices report the extent to which the actual data fit the theoretical expectations created by the many-facet Rasch model. Outfit statistics are unweighted and thus can be influenced more by extreme values and outliers. On the other hand, infit statistics are weighted and are therefore more often used by Rasch researchers in reporting model fit. All fit statistics are positive values. In the field of social sciences, in recognizing that human behaviors do normally vary, it is conventionally accepted that no model will fit the data perfectly, nor is it desirable to find a model perfectly matched by the data. If the fit indices are too low, it means that the data lack variability and this situation would be characterized as *overfit*. If the fit indices are too high, it means that the data behave more erratically than systematically, and thus the situation would be characterized as *underfit*. A rule of thumb for determining the acceptable range of fit statistics is to use the range within 95% confidence interval from the mean fit indices, which can be calculated by two standard

deviations (*SD*) above and below the mean of infit statistics (i.e., $M \pm 2(SD)$). There are other researchers who suggested more conservative criteria, e.g., Lunz, Wright and Linacre (1990), who proposed 0.5 – 1.5 as the acceptable range of fit indices, or McNamara (1996), who suggested an even stricter range of 0.8 – 1.2 for standard language test items. However, these guidelines are advised based on well-behaved data from multiple-choice questions, which may be an unrealistic fit goal for "messy" performance rating data (Bonk & Ockey, 2003, p. 96). Given the nature of the test being a performance assessment, this study adopts the 95% confidence interval method for calculating the acceptable range for the fit indices. Under this guideline, our rater fit statistics should fall between 0.06 and 2.02 (i.e., $1.04 \pm (2 \times 0.49)$). This means that all of the six raters are within acceptable fit statistics, except for Rater 2 who appeared to be on the edge of being underfit given Rater 2's infit mean square of 1.83. On the lower end of our fit spectrum, Raters 1 and 3 appeared slightly overfit, however, still within an acceptable range as well. With the information from the raters' fit statistics, we can further distinguish between Rater 2 and Rater 5; despite their equal logit at -0.75, Rater 5 was much more consistent than Rater 2 in using the rubric.

Student measurement report. The logit values for the student facet represent the ability these students displayed in their roleplay performance. The students' logits occupy the range between -0.51 logit to 1.61 logit, making up a range of 2.12 logits for the whole group. FACETS reported separation reliability of 0.91, which for student measurement report is analogous to Cronbach's alpha reliability estimate, indicating that the current test procedure is quite reliable in separating the students based on their different levels of ability on the construct we are measuring. This is also confirmed by the significant fixed (all same) chi-square (χ^2=1474.9, *df*=147, p<.00) which rejected the null hypothesis for no difference in their ability levels. The separation index of 3.11 further suggests that there are about three statistically distinct levels of

student ability in this dataset, meaning that this test was able to statistically categorize students into three groups—hypothetically, the very high, middle, and very low performers.

Table 7.7
Selected Measurement Report for Underfit Students

Student	Ability Logits	Model *S.E.*	Infit MnSq	Outfit MnSq
…				
ID 33	0.44	0.13	1.79	1.76
ID 48	0.96	0.14	1.80	1.84
ID 59	1.23	0.16	1.76	1.50
ID 66	1.16	0.15	1.90	2.51
ID 83	0.23	0.12	1.68	1.61
ID 167	1.18	0.16	1.79	1.24
…				
M	0.49	0.13	1.03	1.04
SD	0.43	0.01	0.31	0.33

Note. Reliability = 0.91; Separation: 3.11; Fixed chi-square: 1474.9 (*df*=147; *p*<.00), RMSEA = 0.13

In terms of fit statistics, the acceptable range calculated from the $M \pm 2(SD)$ formula gave us a range between 0.42 and 1.66 for acceptable student fit behaviors. Following this guideline, there was no student overfit, but there were six underfitting students whose infit statistics were above this upper control limit. Another model fit indicator reported here is the Root Mean Square Error of Approximation (RMSEA) which estimated the lack of fit in the model compared to the idealized perfect fit scenario (Tebachnick & Fidell, 2013). Ideally, we want the RMSEA to be as low as possible. A good-fitting model should have the RMSEA below 0.06, and the upper limit for acceptable fit is 0.10 (Tebachnick & Fidell, 2013, p. 722). The current model in this study reported the RMSEA at 0.13 which indicated that there is a slightly exceeding amount of unexpected observations for student ability estimation. According to the infit statistics, we identified six underfitting students. Table 7.7 presents the measurement report of those underfitting student test takers. The full version of the student measurement report is provided in Appendix B. The underfitting students all have logit scores higher than 0.20, which falls within the logit range of ability in which there is only one item with the level of difficulty to match. This further shows that this rubric and test task may be more reliable to assess lower ability students, and more challenging items should be added to further improve the test reliability.

Item measurement report. In interpreting the item logits, higher logit values indicate that the items were more difficult for this group of student participants, while lower logit values indicate otherwise. The detailed results from FACETS of item facet made up of the eight items in the roleplay are summarized in Table 7.8.

Table 7.8
Measurement report for items

Item	Difficulty Logits	Model *S.E.*	Infit MnSq	Outfit MnSq
Self-Introduction (SI)	-0.51	0.04	1.02	1.03
Work Talk (WT)	-0.13	0.03	1.00	1.03
Contact Exchange (CE)	0.98	0.02	1.09	1.07
Post-conference Arrangement (PA)	0.18	0.03	1.15	1.20
Activity Termination (AT)	0.16	0.03	1.00	1.00
Understanding (U)	-0.47	0.03	0.82	0.80
Alignment (AL)	-0.26	0.03	0.96	1.13
Affiliation (AF)	0.05	0.03	1.04	1.10
M	0.00	0.03	1.01	1.04
SD	0.48	0.00	0.10	0.12

Note. Reliability = 1.00; Separation: 15.41; Fixed chi-square: 2111.5 (*df*=7; *p*<.00), RMSEA = 0.03

Based on the difficulty logits, the most difficult activity was the *Contact Exchange* (CE) with the highest logit of 0.98. The easiest activity is *Self-Introduction* (SI) with the lowest logit of -0.51. Given the range of student ability in the study was between -1.00 to 1.11, the majority of the students were likely to have successfully managed their SIs, while most students except the very top ones had less than a 50 percent chance of successfully carrying out CE.

If we only consider the production actions, the item SI was the easiest for the students to manage, followed by Work Talk (WT), Activity Termination (AT), Post-conference Arrangement (PA) and finally CE. For the recipient actions, the item Understanding (U) was the easiest recipient action to manage. Following item U, students had a harder time managing their Alignment (AL), and Affiliation (AF) respectively. Upon inspecting the item measurement report, while it appeared on the vertical ruler that SI and U shared a rank for being the easiest items of all the activities, we can see that the difficulty logits of U (-0.47) is marginally higher than SI (-0.51), meaning that SI is the easiest item for this group of student test takers overall.

The separation index was reported at 15.41, suggesting that the activities are reliably different in terms of their difficulty. The fixed chi-square value is also significant (χ^2=2111.0, *df*=7, *p*<.00), which confirmed the earlier finding. Regarding the amount of error, all activities have a very small amount of error, and the Cronbach's alpha of 1.00 for their reliability was very high. This provides a credible indicator that all these activities elicited different aspects of interactional competence (IC) and that the raters were able to consistently rate student performances on these activities independently and reliably across different actions.

The acceptable range of fit statistics calculated from the *M*±2(*SD*) formula gave us a range between 0.81 and 1.21 for acceptable item fit. None of the items on the rubric display erratic behavior as their fit indices mostly centered around one. Normal fit statistics for the items facet can also be used as an indicator that they all form a single construct under the Rasch model (Bond & Fox, 2015), and this further provides support for unidimensionality assumption required for FACETS analysis.

Category measurement report. For the rating scale that this study utilized in assessing aspects of interactional competence represented in each item, FACETS also provides a category measurement report, which offers a diagnosis into the extent to which each of the six score categories – from the lowest score of zero to the full score of five – were optimally performing. A well-functioning rating scale would create clear hierarchical patterns of student abilities for each scale step, which allow for clear differentiation of abilities for each score level (Myford & Wolfe, 2004). The partial credit scale statistics presenting the examinations into each item are displayed in Table 7.9. Columns 3-5 under *Data* show the frequencies and percentage of each assigned score used by all the raters in the dataset. The quality control columns compares the average student ability logits in each category from the data (column 6) to the expected value of the ability measure if the data fit the Rasch model (column 7), and provide the fit statistic in the

form of outfit mean square (column 8) to display the extent to which the data fit the Rasch model. The infit statistics are not reported because they are approximately the same as the outfit statistics in this case. To interpret the fit statistics, Linacre (2010) advises that the outfit square values closer to 1.0 indicate a reasonable level of fit, while values much larger than this suggest unexpected or erratic observations. Also, because Rasch theory states that the advancing categories in a rating scale resemble the higher level of ability on the construct, we should observe an increasing pattern on the average measures further up the scale levels, and their values should be close to the model's expected measures.

For step calibrations of the rating scales, in column 9, FACETS provides Rasch-Andrich thresholds estimating the ability logits where the two adjacent categories are equally likely (Linacre, 2010). Along with the Rasch-Andrich threshold estimates, column 10 displays their corresponding standard errors for the step calibrations. While the average measures represent the central tendency of the ability logits in each step, the thresholds represent the boundaries between score levels – the locations on the ability logit scale where the probability of being assigned score *n* was surpassed by the probability of being assigned score *n+1*. Similar to the way average measures are interpreted, to make inferences that each category level represents an increasing ability on the construct, the Rasch-Andrich threshold measures should also display an advancing pattern going up the score levels.

Table 7.9
Rating Scale (Partial Credit) Statistics

		Data			Quality Control			Rasch-Andrich Thresholds	
Item	Score	Category Counts	%	Cum.%	Avge Meas	Exp. Meas	Outfit MnSq	Measure	Error
SI	0	3	0%	0%	0.48	0.39	1.1		
	1	20	2%	3%	0.80	0.53	1.3	-1.44	0.58
	2	83	9%	12%	0.71*	0.7	1.0	-0.81	0.22
	3	325	37%	49%	0.85	0.88	0.9	-0.58	0.11
	4	288	32%	81%	1.10	1.08	1.0	1.10	0.07
	5	169	19%	100%	1.29	1.29	1.0	1.72	0.09
WT	0	20	2%	2%	0.28	0.03	1.4		
	1	12	1%	4%	0.12*	0.17	0.9	0.61	0.23
	2	83	9%	13%	0.34	0.33	1.0	-1.69	0.19
	3	286	32%	45%	0.47	0.51	0.9	-0.82	0.11
	4	313	35%	80%	0.66	0.70	1.0	0.51	0.07
	5	174	20%	100%	1.00	0.91	0.9	1.39	0.09
CE	0	358	40%	40%	-0.75	-0.77	1.2		
	1	88	10%	50%	-0.60	-0.61	1.3	0.71	0.08
	2	113	13%	63%	-0.43	-0.45	0.8	-0.78	0.08
	3	160	18%	81%	-0.31	-0.28	1.1	-0.71	0.08
	4	111	13%	93%	-0.11	-0.12	1.0	0.17	0.10
	5	58	7%	100%	-0.05	0.04	1.0	0.61	0.14
PA	0	60	7%	7%	0.14	-0.21	1.8		
	1	30	3%	10%	-0.02*	-0.07	1.1	0.55	0.14
	2	115	13%	23%	0.08	0.09	1.0	-1.34	0.12
	3	292	33%	56%	0.16	0.26	1.1	-0.76	0.09
	4	237	27%	83%	0.45	0.45	0.9	0.56	0.07
	5	154	17%	100%	0.69	0.64	0.9	0.98	0.09
AT	0	27	3%	3%	0.03	-0.15	1.2		
	1	108	12%	15%	0.11	0.00	1.2	-1.46	0.20
	2	190	21%	37%	0.08*	0.17	0.7	-0.48	0.10
	3	309	35%	71%	0.30	0.35	1.2	-0.23	0.08
	4	149	17%	88%	0.58	0.54	0.8	1.18	0.08
	5	105	12%	100%	0.83	0.73	0.9	0.99	0.11
U	0	15	2%	2%	0.60	0.31	1.6		
	1	24	3%	4%	0.12*	0.44	0.5	-0.09	0.27
	2	37	4%	9%	0.50	0.58	0.7	0.08	0.17
	3	142	16%	25%	0.61	0.75	0.6	-0.68	0.13
	4	215	24%	49%	0.87	0.93	0.7	0.42	0.08
	5	455	51%	100%	1.22	1.13	0.9	0.28	0.07
AL	0	19	2%	2%	0.83	0.15	3.0		
	1	38	4%	6%	0.21*	0.29	1.0	-0.47	0.24
	2	54	6%	13%	0.34	0.44	0.7	0.01	0.15
	3	203	23%	35%	0.47	0.61	0.7	-0.8	0.11
	4	293	33%	68%	0.83	0.79	0.7	0.33	0.08
	5	281	32%	100%	1.05	0.99	0.9	0.93	0.08
AF	0	54	6%	6%	0.26	-0.09	1.9		
	1	52	6%	12%	0.04*	0.05	1.2	0.02	0.15
	2	117	13%	25%	0.06	0.20	0.6	-0.69	0.11
	3	176	20%	45%	0.26	0.36	0.7	-0.13	0.09
	4	256	29%	74%	0.53	0.54	0.8	0.08	0.07
	5	233	26%	100%	0.81	0.73	0.9	0.73	0.08

Note. The asterisks (*) in average measures indicate unexpected observations where the measures are not advancing from its lower adjacent category.

According to the rating scale quality controls, multiple issues were found in the score categories 0, 1, and 2 in most items. For example, in item SI, the average measure of category 1 (0.8) was observed to be higher than that of category 2 (0.71), and much higher than the model's expected measure (0.53). This unexpected behavior was also revealed in its fit statistics of 1.3

which is considerably higher than the advised upper limit of 1. The category probability curves for each item shown in Figure 7.3 provide a visual representation which further corroborates the observed issues as there are not any discernible peaks among the probabilistic curves for categories 0, 1, and 2, suggesting that most of the time raters have a difficult time distinguishing students' performance between these score levels.

Inspecting the category probability curves (Figure 7.3) also revealed troubles for categories 3 and 4 in all items given their lack of distinct peaks. Myford and Wolfe (2004) noted that for a well-functioning rating scale, each category should have a separate peak which represents a sizable range in which that category is likely to be used. It should be noted that this does not pose a major problem to the overall model (Linacre, 2010). However, Linacre suggested that in the case that a smooth increment pattern of advancing abilities is needed, adjustments can be made by combining categories to dispose of the ones that are not functioning well. This may be done by collapsing categories with closely positioned Rasch-Andrich thresholds or combining categories with disordered average measures.

Given the results from the category measurement report, the study has experimented with two modifications to the rating scale. First, categories 0, 1, and 2 are combined as these categories tended to be statistically indistinguishable, and second, category 3 is to be merged with category 4 to improve the category fit between the data and the Rasch model in future versions of the rubric.

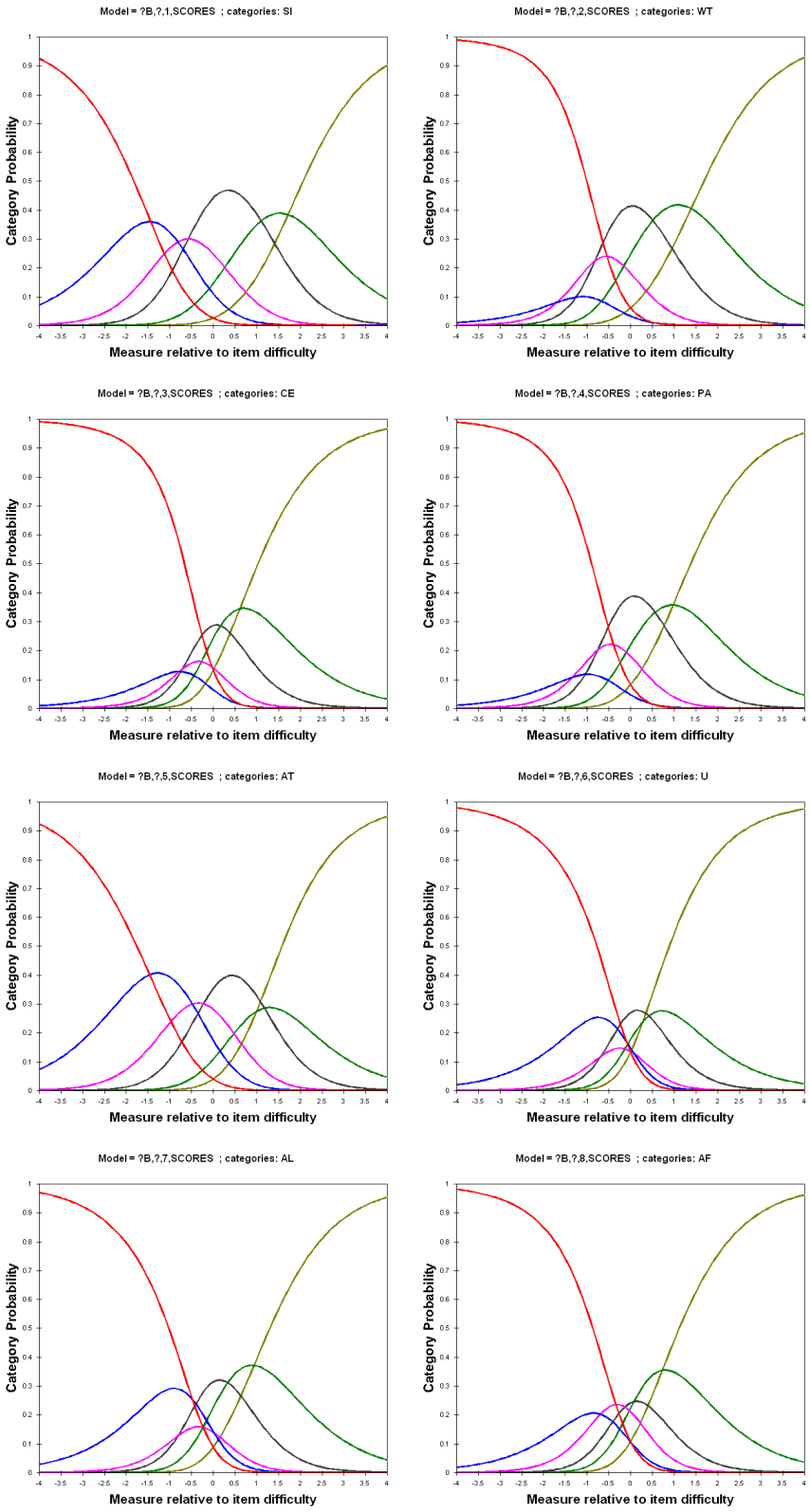


Figure 7.3 *Rating Scale Category Probability Curves*

Modified category measurement report. According to the modifications outlined above, the data have been temporarily recoded to explore whether this change could improve the rating scale. Instead of the six-level scale of 0-5, the modified rating scale only consists of three levels on a scale of 1-3. Categories 0, 1, and 2 were transformed to 1; categories 3 and 4 were transformed to 2, and the full score of 5 was adjusted to the full score of 3. Table 7.10 reports the category statistics of the modified rating scale, and Figure 7.4 shows the category probability curves after applying this transformation.

Table 7.10
Modified (3 Levels) Rating Scale (Partial Credit) Statistics

Item	Score	Data			Quality Control			Rasch-Andrich Thresholds	
		Category Counts	%	Cum. %	Avge Meas	Exp. Meas	Outfit MnSq	Measure	Error
SI	1	106	12%	12%	-0.34	-0.52	1.1		
	2	613	69%	81%	0.21	0.24	1.1	-1.9	0.11
	3	169	19%	100%	0.96	0.97	1	1.9	0.09
WT	1	115	13%	13%	-0.38	-0.54	1.1		
	2	599	67%	80%	0.14	0.21	1	-1.82	0.11
	3	174	20%	100%	1.07	0.94	0.9	1.82	0.09
CE	1	559	63%	63%	-1.83	-1.84	1.1		
	2	271	31%	93%	-1.11	-1.11	0.9	-0.75	0.08
	3	58	7%	100%	-0.65	-0.49	1.1	0.75	0.14
PA	1	205	23%	23%	-0.59	-0.86	1.2		
	2	529	60%	83%	-0.26	-0.14	1.2	-1.45	0.09
	3	154	17%	100%	0.61	0.56	1	1.45	0.1
AT	1	325	37%	37%	-1.23	-1.28	1		
	2	458	52%	88%	-0.62	-0.56	1.2	-1.26	0.08
	3	105	12%	100%	0.27	0.11	0.9	1.26	0.11
U	1	76	9%	9%	0.04	0.2	0.9		
	2	357	40%	49%	0.76	0.89	0.7	-1	0.13
	3	455	51%	100%	1.73	1.6	0.8	1	0.07
AL	1	111	13%	13%	-0.16	-0.25	1.2		
	2	496	56%	68%	0.43	0.46	1	-1.39	0.11
	3	281	32%	100%	1.2	1.17	0.9	1.39	0.08
AF	1	223	25%	25%	-0.62	-0.66	1.1		
	2	432	49%	74%	-0.05	0.03	0.9	-0.98	0.09
	3	233	26%	100%	0.78	0.69	0.9	0.98	0.08

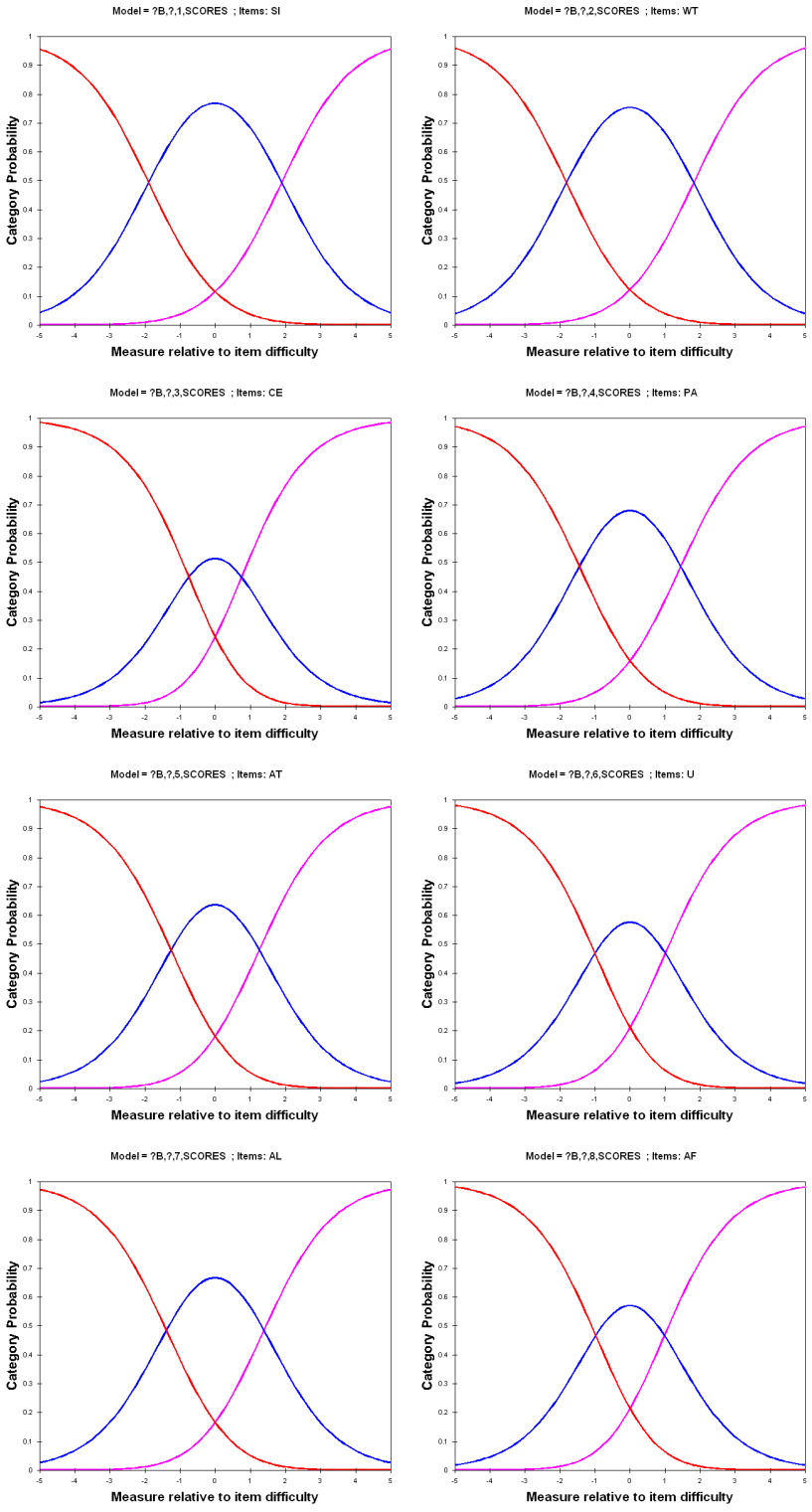


Figure 7.4 *Modified (3 levels) Rating Scale Category Probability Curves*

There are several noticeable improvements in the quality control columns which show the average measures advancing as expected for all items, and that the average measures are now closer in value to the expected measures with acceptable category fit. The Rasch-Andrich thresholds are also advancing monotonically, and the error of the threshold estimations are lower as well. The category probability curves of the modified rating scale (Figure 7.4) also show improvement as distinct peaks for all three categories are observed in all the items.

FACETS Interaction Analysis

FACETS analysis provides bias analysis reporting the interactions between various facets that we specified in the model. FACETS treats the bias/interaction analysis as a secondary analysis which it performed on the main model (Linacre 2010), testing the null hypothesis that there is no discernable pattern between any two facets beyond the error that the model already expects. To explore all possible biases, this study looked into the interactions between raters and items, raters and students, as well as the interaction between items and students. The results of all the bias-interaction analyses are discussed below in that order.

Rater-item interaction. Generating results for rater-item biases, FACETS analysis provides a pair-wise test of whether the raters were behaving similarly across all items. In any rater-item pairings in which the interactions were significant, it is then interpretable that the rater in focus had been variably harsher or more lenient in rating that item beyond what could be expected by the Rasch model. In this study, there were 48 possible cases of interactions (six raters × eight items), and the bias analysis revealed that 33 of them, roughly up to 68% of these interactions, were significantly biased ($df = 147$, $p<.05$).

Table 7.11 presents detailed results of the bias analysis report on these 33 significant interactions. To provide the baseline for comparison, the first two columns show the item ID and

the overall item difficulty (based on ratings from all six raters) expressed in logits, and the paired rater ID and their severity logits (based on their rating across eight items) are reported in the next two columns. The next three columns provide a comparison in the unit of raw scores between the observed scores and the model generated expected scores without the bias consideration. Given the possible rating between 0 and 5 for 148 students, the maximum observed score for each rater on a single item is 740 (5×148). Column five reports the observed aggregated scores for the rater-item pairs; column six presents the predicted scores from the Rasch model; and column seven displays the difference between observed and expected scores averaged to present the amount of bias per student. From these three columns, we can begin to see, for example, that Rater 2 was particularly more lenient in rating SI and awarded the student 0.65 points higher on this item on average. On the other hand, Rater 3 was harsher in rating SI and on average awarded the students 0.47 lower than what we would expect.

Column eight presents the bias size indicating both the degree and direction of the bias in logit unit, and column nine presents the standard error of the bias estimates. Positive bias size signifies that the rater rated the item more leniently than expected. Negative bias size signifies that the rater was harsher than expected in rating the item. The next two columns (columns 10 and 11) display *t*-statistics and its probability (*p*) showing the statistical significance of the null hypothesis. When there are more than 30 observed cases, *t*-statistics is normally distributed and can be interpreted similarly to the *z*-scores. In those cases, significant biases are those with *p* smaller than 0.05 and *t*-score greater than ±2, suggesting that, with more than 95% certainty, the null hypothesis that there is no bias should be rejected.

The last column presents a fit statistic which indicates the amount of the overall misfits that remained after the bias is accounted for. It should be noted that this fit statistic does not report that fit of the bias terms; thus, the range of acceptable values for the infit mean square is

interpreted differently than the ones in measurement reports (Linacre, 2018). The mean square values which are less than one indicate that these biases can be used to explain the misfits, while the mean square values beyond one suggest that the misfit in the data are there due to other unknown sources.

Table 7.11
Bias Calibration Report: Rater-Item Interaction

Item ID	Item Difficulty (Logit)	Rater ID	Rater Severity (Logit)	Observd Score	Expctd Score	Obs-Exp Average	Bias Size (Logit)	Error	*t*	*p*	Infit MnSq
SI	-0.51	Rater2	-0.26	654	558.00	0.65	1.01	0.12	8.62	0.000	1.0
SI	-0.51	Rater3	-0.03	461	530.72	-0.47	-0.54	0.08	-6.31	0.000	0.5
SI	-0.51	Rater4	0.36	440	481.09	-0.28	-0.30	0.08	-3.55	0.001	1.0
SI	-0.51	Rater5	-0.25	528	557.48	-0.20	-0.25	0.09	-2.75	0.007	0.9
SI	-0.51	Rater6	-0.02	559	529.81	0.20	0.25	0.09	2.64	0.009	**1.2**
WT	-0.13	Rater2	-0.26	605	562.39	0.29	0.39	0.10	3.87	0.000	**1.2**
WT	-0.13	Rater3	-0.03	458	532.63	-0.50	-0.47	0.07	-6.34	0.000	0.5
CE	0.98	Rater2	-0.26	282	331.13	-0.33	-0.16	0.06	-2.76	0.007	**1.4**
CE	0.98	Rater4	0.36	211	155.91	0.37	0.22	0.06	3.61	0.000	0.9
CE	0.98	Rater5	-0.25	289	329.71	-0.28	-0.13	0.06	-2.29	0.024	0.9
PA	0.18	Rater1	0.21	481	434.68	0.31	0.22	0.07	3.07	0.003	0.3
PA	0.18	Rater2	-0.26	565	526.85	0.26	0.24	0.08	2.93	0.004	**2.8**
PA	0.18	Rater3	-0.03	446	484.67	-0.26	-0.18	0.07	-2.75	0.007	0.5
PA	0.18	Rater4	0.36	439	398.70	0.27	0.17	0.07	2.58	0.011	**1.2**
PA	0.18	Rater5	-0.25	498	526.08	-0.19	-0.16	0.07	-2.15	0.033	0.9
PA	0.18	Rater6	-0.02	425	483.20	-0.39	-0.27	0.07	-4.11	0.000	**1.3**
AT	0.16	Rater1	0.21	414	384.74	0.20	0.16	0.07	2.14	0.034	0.5
AT	0.16	Rater3	-0.03	394	428.41	-0.23	-0.19	0.07	-2.53	0.012	0.6
AT	0.16	Rater5	-0.25	436	469.40	-0.23	-0.18	0.07	-2.50	0.014	**1.1**
AT	0.16	Rater6	-0.02	473	427.05	0.31	0.25	0.08	3.38	0.001	**1.3**
U	-0.47	Rater1	0.21	521	578.68	-0.39	-0.26	0.06	-4.10	0.000	0.3
U	-0.47	Rater2	-0.26	711	648.86	0.42	0.96	0.17	5.55	0.000	0.9
U	-0.47	Rater3	-0.03	712	618.98	0.63	1.22	0.18	6.93	0.000	0.5
U	-0.47	Rater4	0.36	466	546.56	-0.54	-0.31	0.06	-5.29	0.000	0.7
U	-0.47	Rater5	-0.25	689	648.35	0.27	0.48	0.13	3.77	0.000	0.8
U	-0.47	Rater6	-0.02	560	617.87	-0.39	-0.32	0.07	-4.67	0.000	**1.1**
AL	-0.26	Rater2	-0.26	510	598.03	-0.59	-0.52	0.07	-7.50	0.000	**1.7**
AL	-0.26	Rater3	-0.03	663	563.86	0.67	0.87	0.12	7.52	0.000	0.7
AL	-0.26	Rater6	-0.02	605	562.64	0.29	0.29	0.09	3.28	0.001	1.0
AF	0.05	Rater1	0.21	412	453.49	-0.28	-0.15	0.06	-2.56	0.011	0.3
AF	0.05	Rater2	-0.26	441	558.87	-0.80	-0.51	0.06	-8.37	0.000	**1.8**
AF	0.05	Rater4	0.36	468	412.37	0.38	0.21	0.06	3.33	0.001	0.8
AF	0.05	Rater5	-0.25	663	558.02	0.71	0.82	0.11	7.32	0.000	0.8

Among the 33 cases of rater-item interactions, items U and PA display significant biases among all six raters; SI displays significant biases among all raters except Rater 1; AT and AF display significant biases among four raters; CE and AL display significant biases among three raters; and WT display significant biases from Rater 2 and Rater 3. Figure 7.5 provides a visual

report of these rater-item interactions in a line graph representing the bias logits of each rater across all items.

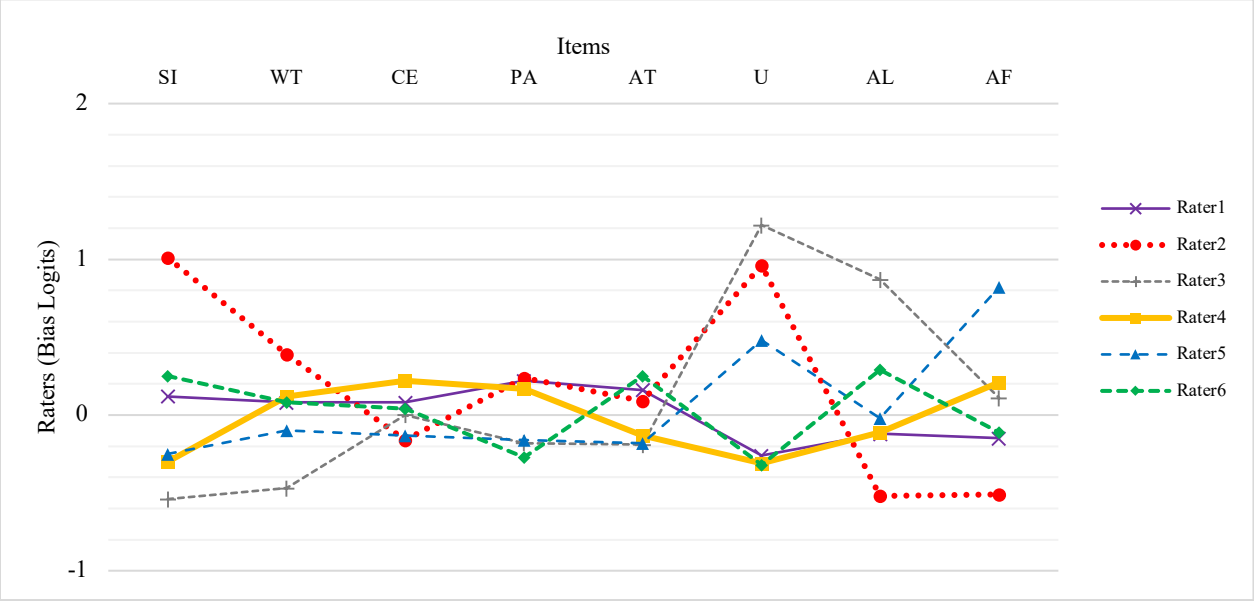


Figure 7.5 *Rater-Item Interaction*

A notable bias pattern which emerged from this interaction analysis is the rater biases on rating U. All raters displayed biases on this item, and the infit statistics for all interactions involving U are around one or lower, indicating that different rater severities on this item can be used to explain the amount of data misfit for U. More specifically, Raters 1, 4, and 6 were harsher than expected in rating this item, while Raters 2, 3, and 5 were, by a large degree, more lenient than expected. Among the raters who appeared to be stricter in rating U, Raters 1 and 4 were those recruited with a background training in conversation analysis. Raters 2 and 3, who were the two most lenient raters for this item, were teachers of the course from which we collected the data. Because this item U was targeted at assessing the students' ability to display understanding appropriately in interaction, it could be that raters with conversation analysis training have higher expectations or notice moments in which students displayed problems managing their understanding displays better than raters without such training.

In another item which was flagged as containing significant biases across all six raters, the item PA also split all the raters into the harsher and the more lenient groups; however, this patter was found to be in a smaller degree compared to item U. In the more lenient group, Raters 1, 2, and 4 all awarded the scores on PA approximately 0.20 logits more lenient that the model expects. The harsher group consists of Raters 3, 5, and 6, who awarded PA about 0.2 logits more severe than the model expects. Again, raters 1 and 4, who were the ones with CA training, behaved similarly regarding their severities in rating PA activity. Conversely, the original raters 2 and 3 this time differed.

From observing the fit statistics in the rater-item interaction, most cases of these rater-item biases have good fit statistics except for those cases which involved Rater 2. While Rater 2 appeared consistently more lenient in rating SI, WT, PA, and U, and consistently more severe in rating CE, AL and AF, the large remaining misfits in most of these items (WT, PA, CE, AL, and AF) suggests that Rater 2's leniency and severity in rating these items do not provide sufficient explanation for the amount of unpredictability in Rater 2's behavior.

So, setting aside Rater 2's bias logits given the high fit statistics of Rater 2's bias estimations, we can see that most of the productive actions (SI, WT, CE, PA, and AT) consistently have smaller bias sizes than the recipient actions (U, AL, AF). In particular, CE, PA and AT all have the bias sizes within ±0.3 logits, while the bias sizes of U, AL and AF are in the range of approximately -0.6 to one logits.

Rater-student interaction. The bias analysis carried out for rater-student interactions identifies whether the raters behaved similarly for all students given their levels of ability. FACETS considers each pairing of raters and students in turn across the ratings of eight items, and from 888 possible biases (148 students × six raters). While most of the interactions are not significant (df=7, p<0.05), the analysis reports 37 significant cases between raters and students or

about 4% significant bias interactions from all 888 cases. Table 7.12 presents a detailed result of the bias analysis on all the 37 significant biases involving five raters and 32 students.

Table 7.12
Bias Calibration Report: Rater-Student Interaction

Rater ID	Rater Severity (Logit)	Student ID	Student Ability (Logit)	Observd Score	Expctd Score	Obs-Exp Average	Bias Size (Logit)	Error	*t*	*p*	Infit MnSq
Rater1	0.21	4	0.71	17	26.75	-1.22	-0.83	0.28	-2.99	0.020	0.2
Rater1	0.21	61	-0.51	21	11.98	1.13	0.70	0.29	2.45	0.044	0.2
Rater2	-0.26	8	0.29	18	27.18	-1.15	-0.79	0.28	-2.85	0.025	**2.0**
Rater2	-0.26	33*	0.44	19	28.69	-1.21	-0.87	0.28	-3.11	0.017	**1.2**
Rater2	-0.26	34	0.82	25	31.92	-0.86	-0.74	0.30	-2.45	0.044	**1.5**
Rater2	-0.26	64	0.27	15	27.03	-1.50	-1.01	0.28	-3.64	0.008	**2.3**
Rater2	-0.26	68	0.33	14	27.63	-1.70	-1.14	0.28	-4.11	0.005	**2.0**
Rater2	-0.26	108	0.71	24	31.06	-0.88	-0.72	0.30	-2.42	0.046	**2.3**
Rater2	-0.26	140	-0.06	32	23.50	1.06	0.89	0.37	2.41	0.047	**1.7**
Rater2	-0.26	178	0.18	17	26.11	-1.14	-0.77	0.28	-2.77	0.028	**1.9**
Rater2	-0.26	179	0.41	19	28.39	-1.17	-0.84	0.28	-2.99	0.020	**1.7**
Rater4	0.36	48*	0.96	20	27.72	-0.96	-0.69	0.28	-2.44	0.045	**3.1**
Rater4	0.36	70	0.54	32	23.22	1.10	0.91	0.37	2.48	0.042	**2.0**
Rater4	0.36	72	0.26	29	19.84	1.15	0.84	0.33	2.55	0.038	1.0
Rater4	0.36	75	0.04	26	17.06	1.12	0.75	0.31	2.44	0.045	0.7
Rater4	0.36	81	0.54	14	23.22	-1.15	-0.73	0.28	-2.63	0.034	0.9
Rater4	0.36	84	0.60	15	23.99	-1.12	-0.72	0.28	-2.61	0.035	**1.4**
Rater4	0.36	86	0.49	14	22.66	-1.08	-0.68	0.28	-2.45	0.044	0.9
Rater4	0.36	105	0.24	30	19.65	1.29	0.96	0.34	2.85	0.025	**1.3**
Rater4	0.36	107	0.30	29	20.40	1.08	0.79	0.33	2.41	0.047	0.8
Rater4	0.36	145	0.96	18	27.72	-1.21	-0.85	0.28	-3.05	0.019	1.0
Rater4	0.36	147	0.92	19	27.31	-1.04	-0.73	0.28	-2.60	0.035	0.9
Rater5	-0.25	15	0.35	36	27.74	1.03	1.21	0.50	2.41	0.047	0.3
Rater5	-0.25	147	0.92	25	32.59	-0.95	-0.83	0.30	-2.75	0.028	0.4
Rater6	-0.02	15	0.35	17	25.33	-1.04	-0.69	0.28	-2.51	0.041	0.7
Rater6	-0.02	64	0.27	34	24.50	1.19	1.10	0.42	2.65	0.033	**2.8**
Rater6	-0.02	83*	0.23	15	24.01	-1.13	-0.72	0.28	-2.62	0.034	**3.2**
Rater6	-0.02	103	-0.19	8	19.00	-1.38	-0.91	0.33	-2.78	0.027	0.9
Rater6	-0.02	104	-0.06	8	20.69	-1.59	-1.04	0.33	-3.19	0.015	0.9
Rater6	-0.02	105	0.24	14	24.17	-1.27	-0.81	0.28	-2.93	0.022	0.5
Rater6	-0.02	106	0.27	12	24.50	-1.56	-1.00	0.29	-3.51	0.010	0.9
Rater6	-0.02	107	0.30	13	24.83	-1.48	-0.95	0.28	-3.39	0.012	1.0
Rater6	-0.02	109	0.60	18	27.99	-1.25	-0.87	0.28	-3.15	0.016	**1.7**
Rater6	-0.02	111	0.73	20	29.15	-1.14	-0.84	0.28	-2.97	0.021	0.5
Rater6	-0.02	160	0.32	17	25.00	-1.00	-0.66	0.28	-2.39	0.048	**1.4**
Rater6	-0.02	161	0.66	21	28.49	-0.94	-0.69	0.29	-2.40	0.047	**2.7**
Rater6	-0.02	162	0.23	14	24.01	-1.25	-0.80	0.28	-2.88	0.024	**1.4**

Note. * = Students who were misfits according to the student measurement report

The bias does not appear to be evenly distributed across raters. Rater 3 displayed no cases of rater-student interaction. Raters 1 and 5 each displayed only two cases of significant bias. Rater 2 displayed nine cases of significant bias. Rater 4 displayed 11 cases of significant bias. Rater 6 displayed 13 cases of significant bias. Fit statistics of these significant biases indicate that the rater-student interaction can account for the misfits in Rater 1 and Rater 5 cases given that their fit indices are smaller than the upper control limit of one (misfits bias fit indices are displayed in

bold format). For Rater 2, who is the most lenient judge according to the model, all the significant biases, except for one, were shown to be the cases graded more severe than what can be anticipated. The higher than upper control limit fit indices for all of Rater 2's significant biases, however, show that these interactions do not account for all the misfits in Rater 2's rating and that there are some other sources which were at play that remain unknown. For Rater 4, the directions of the bias are mixed, and apart from Student#48 whose overall performance was flagged as being too noisy, most of Rater 4's misfits can be explained by this interaction with student ability. Lastly, for Rater 6 who displayed the most interactions with student ability, almost all cases of significant bias are when the ratings were harsher than what the model expected. Interestingly, among the 13 cases of significant bias, five students who received significantly more severe ratings from Rater 6 (Students#103-107) were student participants of the same group performance. While Rater 4 also displayed some bias towards Students#105 and #107, the bias from Rater 4 was in a more lenient direction and did not involve the performance at the group level. This provided us with evidence that some raters may be influenced by the group interaction more than other raters, especially in the negative direction. To some raters, the overall incompetent performance at the group level can negatively bias the rater towards giving all the members lower ratings despite the differing individual ability.

Considering the proportion of significant bias at about 4%, these rater-student interactions are small. This could be due to some idiosyncratic features each student displayed in their interactional performances which unexpectedly influenced the raters' judgment. However, the information gleaned from rater-student interactions can reveal raters' individual tendency of interactions when rating students at different ability levels. For raters 2, 4, and 6, who displayed a larger portion of significant rater-student biases, Figures 7.6-7.8 present line graphs comparing student's overall ability logits and his or her absolute logits when the biases from each rater were

taken into account. In these line graphs, student logits were arranged from the lowest ability to the highest ability on the other end of the scale. The solid lines display the students' overall ability logits, and the dash lines show their absolute measures (their overall logits + bias logits). When the absolute measures fall under the student ability lines, this means that the rater had rated those students more severely than the Rasch model expected and vice versa.

Looking into these tendencies, we begin to see that Rater 2 (Figure 7.6) generally rated students with lower ability more severely up to about 0.5 logits. Beyond that tipping point, Rater 2 displayed a tendency to be more lenient towards more highly capable students. Rater 4 (Figure 7.7) displayed a reversed tendency to Rater 2 in that lower ability students would receive more lenient ratings up to about 0.5 logits, where the ratings became generally stricter. Lastly, Rater 6 (Figure 7.8) displayed biases at both extreme ends of the ability levels. For students approximately lower than zero ability logits, Rater 6 almost exclusively behaved more severely. In contrast, Rater 6 was seen to be much more lenient for students whose ability levels were 1 logit and above. The fact that Rater 4 was much more trained in conversation analysis might be a relevant factor in flipping this tendency we observed in Raters 2 and 6, for CA training could have equipped Rater 4 with finer-grained observations allowing Rater 4 to distinguish between students higher up in the rank, and at the same time, to recognize and award nuanced accomplishments among lower ability students.

Figure 7.6 *Rater-Student Interaction Analysis (Rater 2)*

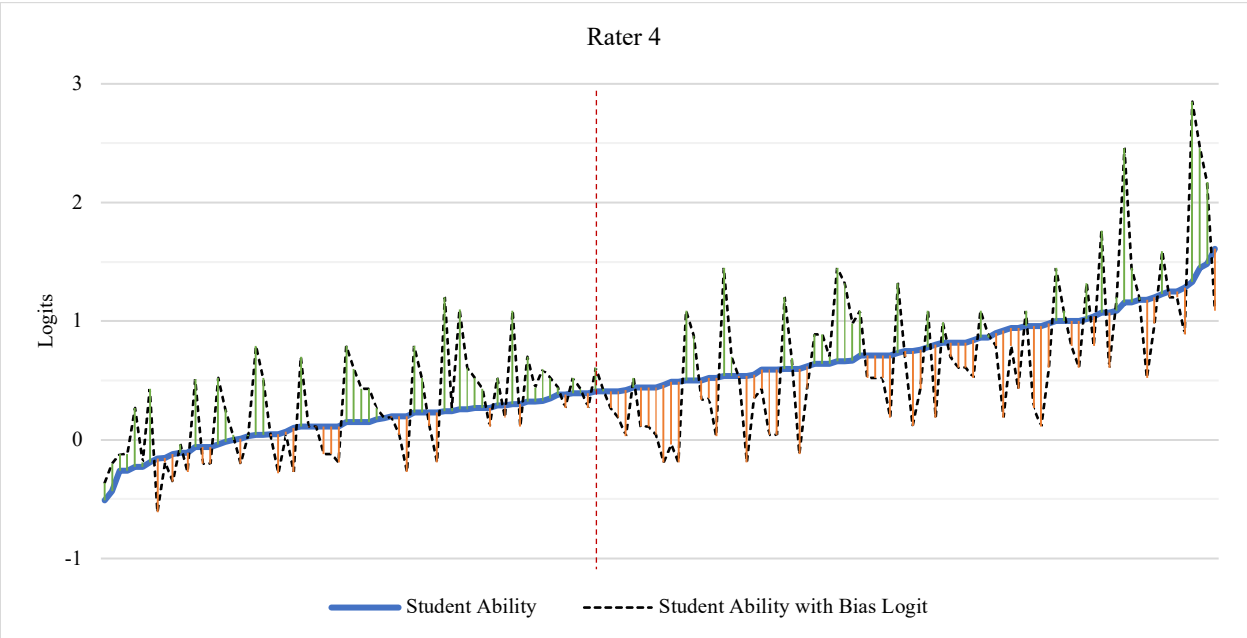


Figure 7.7 *Rater-Student Interaction Analysis (Rater 4)*

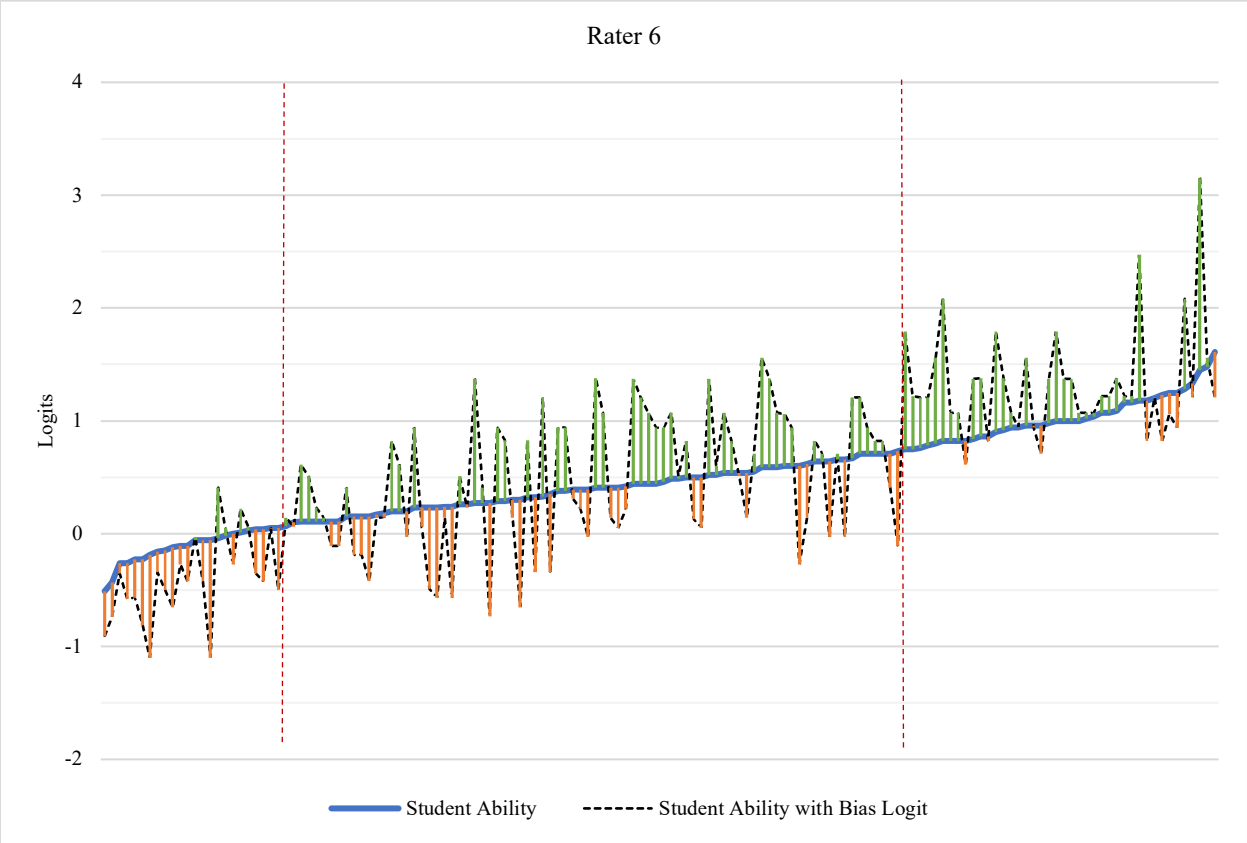


Figure 7.8 *Rater-Student Interaction Analysis (Rater 6)*

Student-item interaction. Finally, a bias analysis was carried out between student and item facets to check whether students behave consistently in all the items, or they behave differently in response to different items. There were 1,148 possible interactions (148 students × eight items) across six raters. The number of student-item interactions with significant bias was 28 cases (df=5, p<0.05) or approximately 3%, involving five items and 27 students. Table 7.13 presents detailed information on the significant student-item interactions.

Table 7.13
Bias Calibration Report: Student-Item Interaction

Student ID	Student Ability (Logit)	Item	Item Difficulty (Logit)	Observd Score	Expctd Score	Obs-Exp Average	Bias Size (Logit)	Error	*t*	*p*	Infit MnSq
111	0.73	SI	-0.51	16.00	22.59	-1.10	-1.21	0.40	-3.04	0.029	**1.1**
122	0.15	SI	-0.51	13.00	19.69	-1.11	-1.08	0.38	-2.87	0.035	**1.1**
22	0.98	WT	-0.13	18.00	23.99	-1.00	-1.04	0.36	-2.92	0.033	**1.8**
30	1.00	WT	-0.13	18.00	24.08	-1.01	-1.06	0.36	-2.98	0.031	**1.7**
81	0.54	WT	-0.13	15.00	21.87	-1.14	-0.94	0.32	-2.95	0.032	0.5
84	0.60	WT	-0.13	16.00	22.22	-1.04	-0.90	0.33	-2.74	0.041	0.6
85	0.80	WT	-0.13	17.00	23.19	-1.03	-0.98	0.34	-2.89	0.034	0.8
86	0.49	WT	-0.13	15.00	21.60	-1.10	-0.89	0.32	-2.79	0.038	0.5
110	1.00	WT	-0.13	15.00	24.08	-1.51	-1.40	0.32	-4.40	0.007	1.0
119	0.05	WT	-0.13	9.00	18.87	-1.65	-1.01	0.30	-3.32	0.021	**1.2**
121	0.07	WT	-0.13	6.00	18.97	-2.16	-1.32	0.33	-3.95	0.011	0.3
15	0.35	CE	0.98	20.00	7.94	2.01	0.90	0.30	2.96	0.032	0.4
17	0.38	CE	0.98	22.00	8.32	2.28	1.07	0.34	3.16	0.025	0.6
27	0.75	CE	0.98	25.00	13.48	1.92	1.14	0.44	2.62	0.047	0.9
32	0.29	CE	0.98	19.00	7.20	1.97	0.87	0.29	2.99	0.031	0.4
42	0.59	CE	0.98	26.00	11.19	2.47	1.51	0.49	3.08	0.027	0.2
47	-0.12	CE	0.98	13.00	3.49	1.58	0.84	0.26	3.19	0.024	0.7
49	1.07	CE	0.98	2.00	17.89	-2.65	-1.49	0.52	-2.85	0.036	1.0
69	1.00	CE	0.98	6.00	17.05	-1.84	-0.82	0.31	-2.64	0.046	**1.7**
83	0.23	CE	0.98	18.00	6.53	1.91	0.85	0.28	3.01	0.030	0.5
115	1.25	CE	0.98	7.00	20.12	-2.19	-0.99	0.30	-3.34	0.021	**1.2**
167*	1.18	CE	0.98	2.00	19.29	-2.88	-1.61	0.52	-3.07	0.028	1.0
31	0.44	PA	0.18	7.00	19.34	-2.06	-1.16	0.32	-3.67	0.015	1.0
32	0.29	PA	0.18	8.00	17.99	-1.67	-0.91	0.31	-2.97	0.031	**1.2**
33*	0.44	PA	0.18	8.00	19.34	-1.89	-1.06	0.31	-3.48	0.018	**1.2**
34	0.82	PA	0.18	13.00	22.09	-1.51	-1.01	0.29	-3.43	0.019	**1.7**
59*	1.23	AT	0.16	12.00	22.49	-1.75	-1.42	0.37	-3.86	0.012	**2.7**
112	0.39	AT	0.16	25.00	16.43	1.43	1.27	0.45	2.84	0.036	0.8

Note. * = Students who were misfits according to the student measurement report

Interestingly, the items which displayed significant amounts of bias are found only in those that targeted the productive side of the interaction. Among these five items, SI and AT displayed the least number of significant bias cases, making the interactions appear situational rather than systematic. Considering SI, for example, Students#111 and #122 unexpectedly found SI items more difficult to complete than other students with the same abilities. However, because the fit statistics of the two cases were slightly more than the upper limit of one, the misfits observed in these students could also be caused by other unknown sources beyond this interaction. The same interpretation could also be applied in PA items. Three interactions out of four which were flagged as significant bias all have higher than the upper limit amount of misfit. With the remaining case, the interaction indicated that Student#31 also had a much tougher time getting a higher score on PA compared to other students with the same ability level. From the current

findings, we cannot draw any substantive conclusions on how these three items (SI, PA, and AT) interact with students at different ability levels. Based on the fit statistics in most of these cases, this could be unrelated to their ability captured in our assessment model. It appears that students' individual characteristics of their performance could have negatively or positively affected their scores on these activities.

According to the number of cases of significant interaction, patterns of bias between students and items which appeared more distinct within this dataset are found in WT and CE. For WT, nine cases of significant bias were found among a small number of students between ability levels of zero and one logits. Because WT was considered an easy item (-0.13 logits), students whose overall ability exceeding -0.13 should have more chances to successfully complete it. The bias analysis has identified significant cases in which WT appeared much more difficult than the model expected for these nine students. Conversely, it could also be interpreted that these students displayed unexpected behaviors in managing their WT which had negatively affected their scores. From this bias analysis, we may note that some higher ability students were still prone to making mistakes in their WT and the prediction of student ability may not indicate the outcome of their interactional performance on WT. This tendency seems to change when students' overall abilities exceed 1.00 logit, in that they would almost always be able to complete WT per all raters' judgments. Figure 7.9 offers a visual comparison between student ability and their ability when WT item bias was taken into consideration.

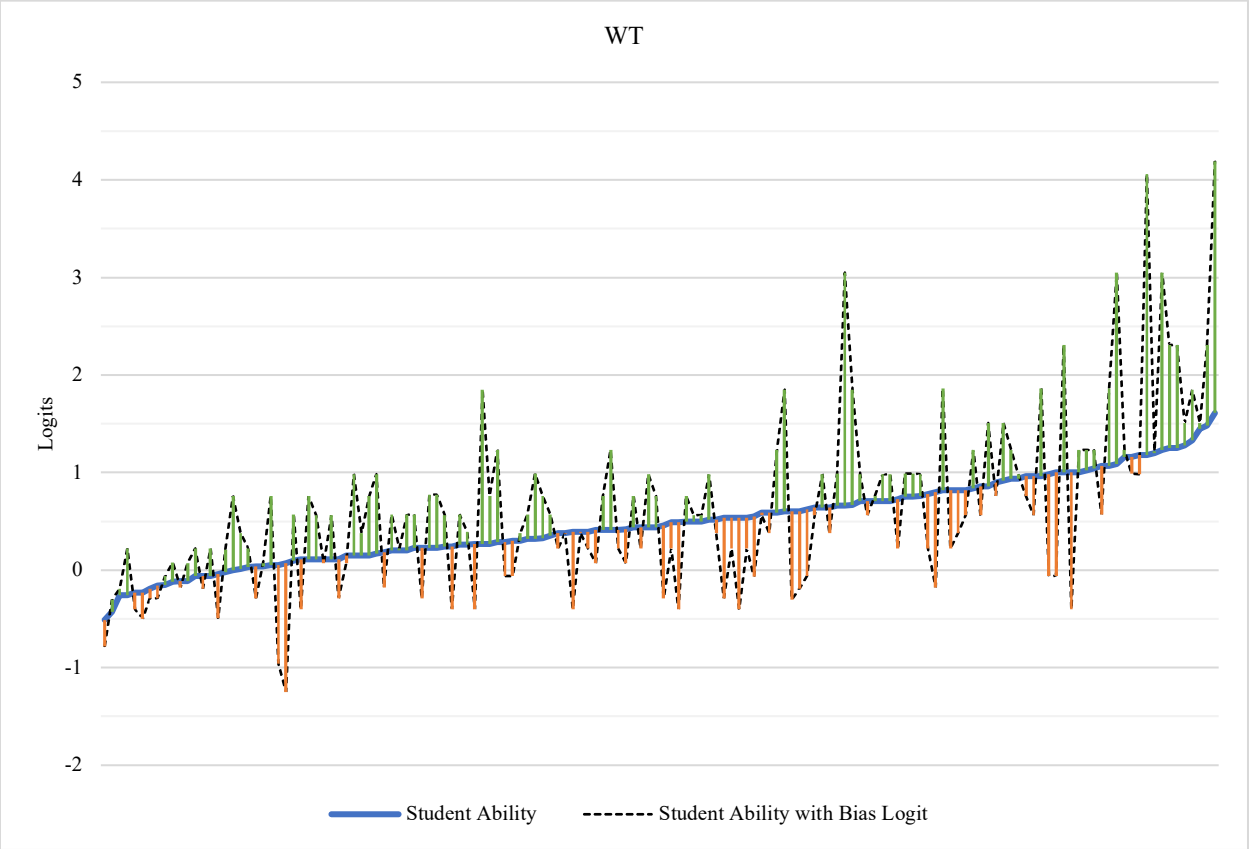


Figure 7.9 *Student-Item Bias / Interaction Analysis - WT*

For CE, which was the most difficult item observed in this dataset, some significant interactions were found mostly among lower ability students who unexpectedly gained a higher score on CE than the model anticipated. To focus our interpretation only on the cases which can be explained by the student-item interactions, we excluded three cases that involved a misfit student (Student#167) and the ones whose fit indices were higher than one (Students#69 and #115). From the 11 significant interactions, we have eight cases of lower ability students who did well beyond the model expectations, and one student who did surprisingly poorly on CE activity. Figure 7.10 presents a visual comparison between student ability and their ability when CE item bias was taken into consideration.

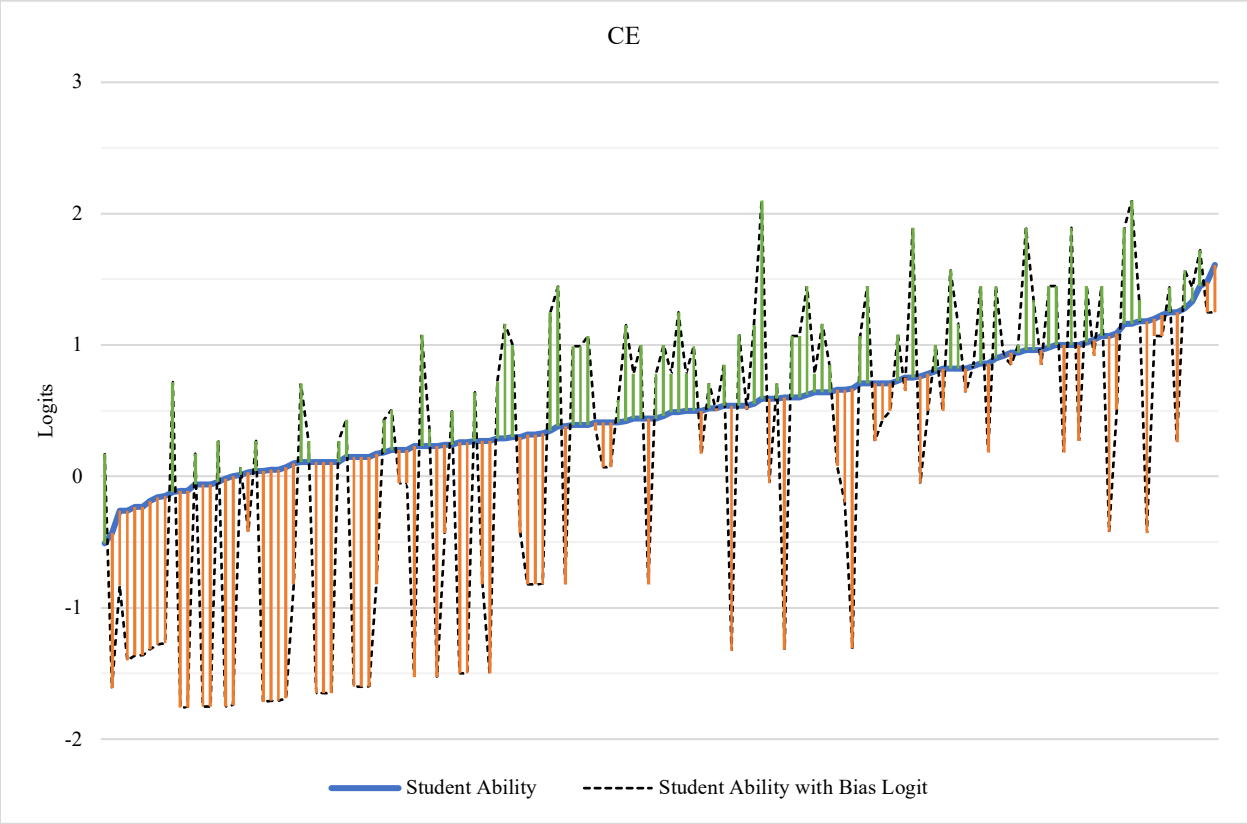


Figure 7.10 *Student-Item Bias / Interaction Analysis – CE*

From the line graph in Figure 7.10, it was clear that lower ability students overwhelmingly found CE much more difficult than other activities required for the task. However, significant biases found among lower ability students may suggest that in some cases, CE can be manageable. The unexpectedness of their success could be due to multiple reasons. Three possible explanations which could have led to such patterns are the saliency of CE activity, a possible peer mediation aiding the students' CE production, and a facility afforded by the roles that they chose to play making it easier for some students to conduct CE. First, given the way the original rubric explicitly required students to exchange contacts and name cards in the roleplay, students may have prepared more for CE compared to other activities. Secondly, a plausible explanation for these interactions could also be that the lower ability students were able to get a higher score than expected because of the help from their peers. Creating one's own opportunity to initiate contact exchange in interaction might be difficult, but it could be much

easier to recognize the activity and continue the CE in the same trajectory after the co-participants had done the necessary interactional work. Also, another possible reason could be attributed to the roles that the students chose to play. Given the design of the task which allows students to choose and prepare for the roles they were going to play during the test by themselves given only that they had to represent a company in attending the hypothetical conference for the roleplay task, there was a wide range of different roles representing various job positions from a diverse field of industries. It might be possible that some roles may inadvertently afford the students an easier to manage CE activity and perhaps also WT activity. Interactions in open roleplays can be unpredictable, and students with a more limited set of skills can also sometimes accomplish the actions.

Conclusion

This chapter presents the findings from quantitative analysis using many-faceted Rasch measurement model. FACETS measurement reports many indicators regarding the reliability of the rating practice as well as the calibrated properties of raters' severity, item difficulty and rating scale category functioning, and the student test takers' reported IC scores. Moreover, FACETS interaction analyses provided an exploratory investigation into potential bias that can affect raters' judgment of IC in different students' ability levels and different interactional activities. To further interpret the results in the light of providing some answers to the study's remaining research questions, the findings from this quantitative analysis will be discussed further in the next chapter.

CHAPTER 8

DISCUSSION OF QUANTITATIVE RESULTS

This chapter discusses the research findings in relation to the quantitative and mixed methods Research Questions 4-6 which the study proposed in Chapter 4. This chapter discusses relevant evidence regarding the validity and reliability of the rating process, exploring the FACETS rater measurement results as well as the bias analyses between the rater facet and the facets representing the student ability and item difficulty based on the *socializing task*. In the final section of the chapter, the evidence taken from qualitative and quantitative analyses regarding the claim that the generated scores represented the students' IC in performing the *socializing task* will be explored, and the strengths of the mixed methods research framework especially for the current study will be discussed.

Research Question 4: *Given the proposed rubric for assessing IC in this roleplay task, how reliable is the rating process in applying the scale to rate the students' performances?*

Rating IC can be an insurmountable task for raters given the level of detail to which they must pay attention (Ross, 2018) while assigning the scores for each targeted activity required by the rubric. To explore the ratability of the IC rubric that the study has proposed, six raters from three different backgrounds were recruited: two came from the original pool of raters who were familiar with the *socializing task* and population of students taking the test, two were ESL teachers working in Hawai‘i, and two were raters with conversation analysis training. All six raters had undergone an individual training with the researcher which lasted two hours to familiarize themselves with the rubric and practice identifying the targeted actions and rating

excerpts of different activities with the researchers before being asked to rate one whole performance to check that they were ready to carry out the ratings on their own.

The six raters who participated in the study showed varying degrees of severity in applying the rating scale during their scoring processes which are commonly reported in rating performance assessments. Although they were not uniform regarding their severity, various quantitative findings from FACETS analysis showed that they were reliable overall. FACETS reliability estimation for the rater facet reported reliability of 0.99 along with normal fit statistics, which indicated that altogether the six of them were optimally self-consistent. Although not to a serious degree, there are some variations among three out of six of the raters' individual fit statistics which may indicate some borderline patterns of misfits. Two raters, Raters 1 and 3, one from the original group of raters and another who was recruited with some conversation analysis (CA) training, displayed a slight overfit behavior, meaning that their ratings may lack variations that are usually expected from rating human performances. One rater, Rater 2, displayed a larger amount of fit statistics close to the upper limit for what would then be considered a much too erratic rating behavior, pointing to the need to provide additional training to improve Rater 2's reliability in using this rubric. The different backgrounds of the six raters, native speaker status, background CA training, and familiarity with the group of test takers, seem to have no influence on how self-consistent they are during the rating process.

Research Question 5: *Given the proposed rubric for assessing IC in this roleplay task, are there any detectable biases in how the raters apply the scale to rate the students' performances?*

Despite the overall reliability reported in FACETS' rater measurement report, when inspecting the interactions between the rater facet and other two facets (item difficulty and student ability), the study showed that raters' severity was in various degrees affected by these

two sources of potential biases. This resulted in the differing severity when rating different items and differing severity when rating students at different ability levels.

Firstly, regarding the rater-item interactions, some raters were found to be inconsistently strict in applying the rubric across different items. For example, the study found that Rater 2, who was one of the more lenient raters overall, was particularly lenient in rating Items SI (Self-introduction activity) and Item U (Understanding display), but was stricter in rating other recipient actions like AL (Alignment display) and AF (Affiliation display). In total, the FACETS interaction study reported 68 percent or 33 significant interactions out of 48 pairings between the six raters on each of the eight items they were rating. Even though the raters' background does not appear to have much influence over their self-consistency in applying the rubric, these different backgrounds may have played a role in how the raters were systematically more or less strict in rating different interactional activities in this roleplay task.

These significant interactions can partly account for the amount of noise identified in the fit statistics reported in the rater measurement study. In general, significant interaction from FACETS bias analysis can offer further explain for the poor fit between the model and the data in that it can in part be caused by the raters being inconsistently strict while rating different items. The bias fit statistics can be used as an indicator of whether this is such the case. When the infit means squares reported in the interaction study are within the recommended degree, that interaction between that rater-item combination can then be interpreted as the reason for that rater's underfitting performance. In this dataset, the fit statistics of most interactions reported to be significant were found to be within the acceptable limit, meaning that their potential inconsistency was mostly caused by the rater-item interactions. There was an exception in the case of Rater 2, who previously had been reported to display the most inconsistent fit statistics which are borderline too high. The fit statistics of the rater-item bias analysis of Rater 2

exceeded the acceptable amount in multiple items. This suggests that Rater 2's inconsistent behavior could not be accounted for by the displayed bias, which we should follow up on in future research. Nevertheless, the findings from rater-item interaction for Rater 2 does provide a guideline for future additional training for Rater 2 in that it needs to focus rating items CE (Contact Exchange), PA (Post-conference Arrangement), AL (Alignment display), and AF (Affiliation display).

Regarding the bias sizes between items targeting productive actions and items targeting recipient actions, the study found that raters' biases were generally larger among the items targeting recipient actions compared to items targeting productive activities. In particular, the report showed that all six raters showed some measure of biases when rating item U, which targeted the display of understanding in interaction throughout the roleplay performance. The bias logits revealed that the six raters could be classified into two groups: three raters were much more lenient than expected when rating item U, and the other three raters were slightly stricter compared to how they rated other items. The two raters with conversation analysis (CA) training background displayed a tendency to rate this item more severely. On the contrary, the raters who were recruited from the pool of original raters were found to be extremely lenient when rating this same item. Because this item U was targeting the students' ability to display understanding appropriately in interaction, it could be that raters with conversation analysis training have a higher expectation or notice moments in which students displayed problems managing their understanding displays better than raters without such training. The finding could also suggest that having CA training can make a difference in the raters' standards on how the recipient's understanding should be displayed. It is premature, however, for the current study to draw any defendable conclusion whether this is the case given that it only has a small sample size of two raters from each background.

Other recipient actions also exhibited larger degrees of biases among the six raters compared to the productive action items, but with a less clear pattern based on their background experiences. Nevertheless, in two out of three items except for AF (Affiliation display), raters with CA training background showed more uniformed rating performances compared to the other four raters.

Overall, the fact that these raters appeared less uniform in their ratings of recipient actions than the productive activities was something that the current study did not anticipate, and it should be addressed immediately in subsequent rater trainings and administrations of the rubric. A potential explanation could be that rating recipient actions of multiple sequences may be more subjected to raters' individual judgments much more than the productive activities which were anchored on more clearly defined sequential positions and expectations.

Another source of raters' bias identified in the FACETS interaction report is the students' ability levels. Some raters were found to display bias when rating students at different ability logits. These biases were not uniformly applicable to all the raters as different raters exhibited different degrees and patterns of biases. While the concerns over rater-student interaction are in a much smaller portion compared to the rater-item interactions given that the reported rater-student interactions are most of the time insignificant, at this explorative stage of developing this test instrument, the raters' bias tendencies gleaned from this interaction report can be helpful for the training of future raters.

Through comparing the student ability logits with their corresponding bias sizes that FACETS reported for each rater, the study found an interesting contrast between one rater with CA training background and other raters without the CA training background. Inspecting raters' bias tendencies when rating the group of students with higher ability logits, raters without CA trainings tended to show a positive bias, meaning that they were more lenient towards this group

of students, while the rater with CA background often displayed a negative bias and was stricter in rating the students higher up the ability levels. This tendency switched when inspecting the bias patterns among students lower down the ability logits scale. The rater with CA background displayed bias in the positive direction, while the raters without CA background became somewhat stricter when rating students with lower ability levels. Future studies need to examine the roleplay performances that motivated such positive or negative biases from raters from different backgrounds to further refine the rating practice of this IC construct.

Research Question 6: *Through mixed methods research, to what extent can the current study argue for the validity of the proposed rubric and rating scale for assessing IC in this context? How do the findings from mixed methods help to strengthen the validity argument?*

Through adopting the sequential mixed methods research framework, conversation analysis (CA) findings which offered detailed descriptions how students carried out their roleplay performances were extracted and assembled into a rubric which focuses on assessing how students display their IC while performing specific actions and activities as part of the roleplay. The rubric proposed in the study contains eight items, and raters were asked to assign a score between 0-5 on these eight items. The scores then have been analyzed by FACETS to reveal not only how reliable the raters were in assigning the scores but also how well the rubric is functioning in assessing the language learners participating in this task.

Through analyzing the results from FACETS item measurement report, we learned that each action/activity was not equally manageable. Granted that the raters were reliable in their score assignment, the contact exchange activity (CE) was by far the most difficult, followed by the talk making post-conference plans (PA) and the activity termination activity (AT) as the top three most demanding interactional activities required by the *socializing task*. Students

reportedly found activities like self-introduction (SI) and talking about their job responsibilities (WT) easier to handle, with SI being the easiest activity to perform, and WT being the fourth easiest after managing appropriate display of understanding (U) and alignment (AL). The reliability estimate in support of this result was very high, and the fit statistics for the item measurement report further indicated that the data and the Rasch model did fit and functioning well.

Moreover, if we cross reference these findings of different item difficulties with the qualitative results from conversation analysis, we can see that SI and WT recruited somewhat similar aspects of IC; both activities required students' ability to manage topics, and WT also demanded students' ability to sustain and develop topics relating to their jobs and work experience. On the other hand, difficult activities like CE, PA, and AT are much more specific. For example, CE required IC in initiating a request, an offer or a suggestion, and PA required managing a suggestion or an invitation sequence.

The fact that these eight actions and activities were all different in terms of their difficulty may indicate that the raters were able to consistently identify these different actions and apply different sets of standards in evaluating each activity separately from one another. This provides reassurance about our identification of these eight activities to represent the IC construct of the *socializing task* given that at the very least, these actions are distinguishable from one another and recognizable by raters as units of observation.

Thus, the corroboration between the quantitative findings and quantitative findings provided some confidence in the validity of the proposed rating scale that it was generating interpretable results and its scores started to gain some meaning.

Another aspect of score validity can be established from the extent to which each scoring step from the scale of 1-5 corresponded with students’ differing levels of effective execution on

each of these activities. Based on the qualitative analysis, the study proposed the 1-5 rating scale steps to represent a successful execution of the activity on one end, and a completely problematic execution on the other. The middle category was devised to capture the kinds of execution that display inadequate control over the sequence organization of the activity.

FACETS student measurement report indicated that the rating practice was adequately reliable in discriminating students at different levels of interactional ability. However, the logit scores, which FACETS produced as the students' ability indicator, were reported to be in the range of -0.51 to 1.61 logits for the whole group of 148 students, making up a narrow range of 2.12 logits. Separation index from student measurement report at 3.11 further suggests that there are only about three statistically distinct levels of student ability which can reliably be captured in the proposed rating scale.

The findings from FACETS category measurement report also suggested that the lower scoring steps of 1-2 were hardly discriminable from the score 0 for the raters on almost all of the eight items. Similarly, the scoring categories 3-4 also showed to overlap considerably on many items as well. The findings from both student measurement report and category measurement report indicated that overall, we should revise the rubric from the 1-5 scale to simply containing three levels of 1-3.

The quantitative analysis discussed above can, therefore, strengthen the validity claim of the proposed rating scale. Following this suggestion to collapsing the scoring steps into only three levels, the proposed rating scale for assessing IC can be further modified to improve the fit between students' differing levels of effective execution and the scoring decisions which raters can make on each of these activities.

In developing and validating a rating scale to measure a task-specific IC such as this one, the mixed methods research design provided a unique angle for the study to gain crucial insights

from the rich description provided by the qualitative analysis and the more decisive and integrative outcomes rendered by quantitative analysis. Not only do the mixed methods approach allow for the quantitative analysis to build off the findings from the qualitative analysis in this sequential mixed methods design, but also the results from both qualitative and quantitative analyses also enhance the interpretations of the findings of one another. This point is evident especially in the way that qualitative and quantitative findings revealed complementing information regarding how the performance of students can be graded. Conversation analysis offered a detailed description of successful and problematic executions in the activity. The quantitative FACETS analysis also took into account the raters' behavior in assigning scores to the group of target learners, and provided a statistically backed indicator that the rating scale can reliably distinguish different levels of successful and non-successful executions as long as the number of scoring levels is not exceeding three. Future revisions of the rubric in assessing these interactional activities can then incorporate the suggested modifications and further improve the validity argument for this proposed rating scale.

CHAPTER 9

CONCLUSION

This concluding chapter provides a summary of the research, discusses the limitations of the study, and considers the implications as well as future possibilities in extending this line of research.

Research Summary

In an effort to develop an assessment instrument in measuring interactional competence (IC) with a method that is congruent with the current research findings on IC and IC development (i.e., Hall et al., 2011; Pekarek Doehler & Pochon-Berger, 2015), the present study investigated students' performances on a multiparty roleplay on a task called *socializing* to explore how students' IC can be validly and reliably identified for assessment purposes. Following the data-driven approach to rubric construction (Fulcher et al., 2011), the study developed a rubric based on the rich descriptions of the students' performance data on the roleplay task. Using the sequential mixed methods design (Greene, J., 2007; Tashakkori & Teddlie, 2003), the study explored empirical evidence garnered through qualitative and quantitative research methods to test if the proposed rubric can provide a valid and reliable measurement of IC on this performance assessment task.

The roleplay data were obtained from 180 university students who took the performance test in groups of four to six students and a total of 34 video-recorded multiparty interactions was included in the qualitative phase of the study. Conversation analysis (Sacks et al., 1974; Schegloff, 2007) was employed to identify comparable interactional activities and determine the interactional methods students utilized in carrying out those activities. Six raters from various

teaching and training backgrounds were recruited to evaluate the students' roleplay performances with the newly proposed rubric developed from the CA findings reported in the qualitative phase of the study. The activities selected to represent the IC construct of the *socializing task* include five specific productive activity and three recipient actions on the overall performance. The productive activities are self-introduction (SI), work talk (WT), business contact exchange (CE), post-conference arrangement talk (PA), and an interaction to bring about the termination of the roleplay performance (AT). Three recipient actions include students' management and display of their understanding (U), students' management of alignment (AL), and finally, their display of affiliative stance (AF). Given the poor camera angle or audio quality of some of the video data, the students' performances that any of the raters were unable to rate were then removed from the analysis, reducing the number of student participants from 180 to 148.

Multifaceted Rasch measurement (Linacre, 1989) with the partial credit scoring model (Masters, 1982) provided integrated measurement reports of the rating practice, which accounted for students' IC ability, item difficulty and rater severity all under the same model. The findings suggested that most raters were reliable in applying the rating scale, but they also differed in their severity in assessing the eight targeted interactional activities. The six raters also demonstrated more uniform rating in assessing productive activities compared to their ratings of recipient actions. Finally, through combine insights from qualitative and quantitative research findings during the process of developing this rubric, the mixed methods research design provided a much-needed framework to exploring the validity evidence of the proposed rubric in assessing IC for the multiparty roleplay performances on the *socializing task*.

Limitations of the Study

Several limitations in this study stemmed from the fact that it has chosen to investigate an already existing performance assessment task instead of designing a new one. While there are several benefits in designing a new task with the specific purpose of assessing IC, i.e., the opportunity to pre-specify the types of interactional situations to target certain interactional phenomena or a more controlled and standardized procedure during the test administration and data collection, the current study chose to investigate an already established task with a claim of assessing social interactional skills in the hope that the approach in developing an IC sensitive rubric that this study implemented could later be applied to other interaction-based tasks that share the same goal.

However, not being involved in the process of designing the task has presented some challenges over the control of several variables which might have been relevant to performance data in this study. First, each of the roleplay performance consisted of an inconsistent number of participants per group, varying from groups of four up to six students depending on the local constraints at the time of the data collection. Though a previous study reported that different group sizes had little effect on their interactional styles (Nakatsuhara, 2011, 2013), having more students in one roleplay performance could make evaluating the performances more demanding for the raters' attention, which could have affected their reliability. It is speculated that this may not impose a significant challenge in evaluating the productive activities since they tend to get brought up sequentially, with students taking their turns one at a time. Nevertheless, evaluating the recipient actions in a larger group of students could be much more demanding for raters given that multiple actions could happen concurrently.

Second, the fact that nearly 90 percent of the student participants of the study were male did raise some concern for how this might have affected the interaction in the performance.

However, given that the population of university students majoring in engineering in Thailand is also mostly male, this gender unevenness between male and female student participants in this study was not really an anomaly.

Another limitation is that the raters were asked to apply the new rubric to rating the performance of the same set of video data from the qualitative analysis. Two possible threats to the trustworthiness of the results include (a) the issue of a possible exaggerated fit between the rubric and this particular set of student performances and (b) the fact that students were graded on a rubric which they did not have a chance to see beforehand, which is pedagogically not ideal. Clearly, future research needs to investigate the implementation of the proposed rubric with a different set of students from the same population to examine if the rubric would produce similar results.

Implications

All in all, aside from all the limitations, the current study has taken another step forward to explore the ways to operationalize IC within an assessment task designed to test L2 social interactions. In the bigger picture, this project has tapped into an interactional construct of as part of any interaction-involved speaking assessments. Part of the movement towards equipping raters with concrete observations in rating scale development especially for the rating scale of IC has seen some notable developments like Youn (2013), who has shown that IC rating criteria, *turn organization* and *engaging with interaction*, can be operationalized to function statistically well in measuring IC. This current study's findings added more evidence that IC can be practically operationalized in interaction-involved speaking assessment to generate statistically reliable observations. The study demonstrated that conversation analysis (CA) can provide such rigorous analytical framework to identify (a) the task-specific interactional phenomena for the IC

assessment construct, (b) the interactional contingency that L2 test takers have to manage in order to effectively implement those interactional phenomena, and (c) the normative desirable interactional achievements that can be used as baseline standards for evaluation purposes. The current study's findings, which suggest that raters who have more CA background training displayed a better fit with our measurement model in general, may warrant the study's implications regarding the rater trainings for any assessing interaction-involved speaking skills in which CA training may also be valuable to improve the reliability of evaluating IC in the future.

While the application of the rating scale proposed in this study outside of the specified context of this study will be limited, the procedures adopted in this study may provide a meaningful frame to address the construct of IC in future studies. Readers should take caution in adopting the full version of the rubric as it may not be directly applicable in assessing IC in other tasks. However, individual items from the rubric which targeted different interactional activities could be more readily transportable for outside application given that the students' performance in those tasks was analyzed to identify if their representative interactional phenomena would validly warrant such application.

Finally, making the rubric descriptors explicit in evaluating IC also has pedagogical implications in support of the movement to integrate CA findings for L2 teaching practices (e.g., Barraja-Rohan, 2011; Huth, 2014; Waring, 2018; Wong & Waring, 2010). A direct implication of fully administering this proposed rubric would be that language learners can get explicit comments and feedback regarding precisely what actions they could manage effectively in a conversation and what actions they still need to improve upon and how. The rubric can be used as resources that teachers can share with their students or use as a means for students' self-evaluations.

Suggestions for Future Study

To seek further evidence for the dependability of the rating practice as well as the reliability and validity of the proposed rubric, future studies could investigate the following research agendas. First, the revised rubric should be implemented, and the results of the new implementation should be analyzed in comparison with the current study's findings. The findings from quantitative analysis have pointed to several revisions that future versions of the rubric should integrate into the current version in order to improve the scoring reliability. This includes (a) collapsing the six scoring steps on the current scale of 0-5 into only three scoring steps from 1-3, (b) discarding item U (displaying understanding) since the item may be too easy for this group of students and it may also be susceptible to rater bias given their different approaches to rating this item, (c) considering bring in more items which can distinguish among upper-level students. Based on the mixed methods findings from cross-referencing the item difficulty and interactional activities' components obtained from CA, candidate items can potentially be selected from actions with a specific sequential structure such as storytelling (Pekarek Doehler & Berger, 2016) or managing disagreement (Pekarek Doehler & Pochon-Berger, 2011).

Moreover, outside of the institutional goal to provide comparable observations of students' IC across the dataset, future studies should also investigate interactional accomplishments other than the ones identified in the selected eight actions and activities that we have included in the proposed rubric. Future research can investigate how students with limited linguistic resources coped with the task requirements. In particular, how did those students mobilize embodied resources to aid their task completion and accomplish fundamental actions in interaction such as pursuing target responses, negotiating turn-taking, or displaying affiliation, etc.? Carrying out these embodied actions successfully can facilitate their roleplay interaction in the direction that is gearing towards completing the task. Although these accomplishments were

not part of the construct that the current rubric is trying to assess, being able to coordinate such actions also demonstrates different aspects of their IC. Therefore, from a test developer's standpoint, such accomplishments also deserve our attention.

To this end, it is crucial that future studies collect video and audio data that allow for such level of fine-grained observations which would facilitate raters to notice necessary nuances of L2 speakers' IC displays. As one of the challenges encountered in this study also stemmed from the poorly-angled video data and of some of the obtained in the first phase of the research which resulted in the loss of 32 participants in later stage of the analysis, it is advisable that future studies take greater care in the process of data collection and, if possible, invest more in the technology that can better capture social interactions as close as possible to what the participants or raters during the roleplay could experience the test performance.

REFERENCES

Al-Gahtani, S., & Roever, C. (2012). Proficiency and sequential organization of L2 requests. *Applied Linguistics, 33*(1), 42-65. doi:10.1093/applin/amr031

Al-Gahtani, S., & Roever, C. (2013). 'Hi doctor, give me handouts': Low proficiency learners and requests. *ELT Journal, 67*(4), 413-424.

Al-Gahtani, S., & Roever, C. (2018). Proficiency and preference organization in second language refusals. *Journal of Pragmatics, 129*, 140-153.

Antaki, C. (2011). Six kinds of applied conversation analysis. In C. Antaki (Ed.), *Applied conversation analysis: Intervention and change in institional talk* (pp. 1-14). Basingstoke, UK: Palgrave Mcmillan.

Atkinson, D. (Ed.) (2011). *Alternative approaches to second language acquisition*. New York: Routledge.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Baker, B. A. (2010). Playing with the stakes: A consideration of an aspect of the social context of a gatekeeping writing assessment. *Assessing Writing, 15*(3), 133-153. doi:http://dx.doi.org/10.1016/j.asw.2010.06.002

Barraja-Rohan, A.-M. (2011). Using conversation analysis in the second language classroom to teach interactional competence. *Language Teaching Research, 15*(4), 479-507. doi:10.1177/1362168811412878

Bejar, I. (1983). *Achievement testing: Recent advances*. Beverly Hills, CA: Sage.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York: Routledge.

Bonk, W., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89-110.

Bonk, W., & Van Moere, A. (2004). *L2 group oral testing: The influence of shyness/outgoingness, match of interlocutors' proficiency level, and gender on individual scores*. Paper presented at the The annual meeting of the Language Testing Research Colloquium (LTRC), Temecula, CA.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing, 26*(3), 341-366. doi:10.1177/0265532209104666

Brouwer, C. E., & Wagner, J. (2004). Developmental issues in second language conversation. *Journal of Applied Linguistics, 1*(1), 29-47.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1-25.

Brown, A. (2005). *Interviewer variability in language proficiency interviews*. Frankfurt: Peter Lang.

Brown, J. D. (2001). Pragmatics tests: Different purposes, different tests. In K. R. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 301-325). Cambridge: Cambridge University Press.

Brown, J. D. (2005). Testing in language programs: A comprehensive guide to English language assessment. New York, NY: McGraw Hill.

Brown, J. D. (2009). Choosing the right number of components or factors in PCA and EFA. *Shiken: JALT Testing & Evaluation SIG Newsletter, 13*(2), 19-23.

Brown, J. D. (2014). *Mixed methods research for TESOL*. Edinburgh: Edinburgh University Press.

Brown, J. D. (Ed.) (2012). *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages*. Honolulu, HI: National Foreign Language Resource Center.

Brown, J. D., & Ahn, R. C. (2011). Variables that affect the dependability of L2 pragmatics tests. *Journal of Pragmatics, 43*, 198-217.

Butler, C. W., Danby, S., & Emmison, M. (2011). Address terms in turn beginnings: Managing disalignment and disaffiliation in telephone counseling. *Research on Language & Social Interaction, 44*(4), 388-358. doi:10.1080/08351813.2011.619311

Button, G., & Lee, J. R. E. (Eds.). (1987). *Talk and social organization*. Philadelphia: Multilingual Matters.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 2-27). London: Longman.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1-47.

Cekaite, A. (2007). A child's development of interactional competence in a Swedish L2 classroom. *The Modern Language Journal, 91*(1), 45-62.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Lingistics, 19*, 254-272.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3-13.

Clayman, S. (2010). Addres terms in the service of other actions: The case of news interview discourse. *Discourse & Communication, 4*(2), 1-22.

Clift, R. (2016). *Conversation analysis*. Cambridge: Cambridge University Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26*(3), 367-396.

Davis, L. (2015). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117-135. doi:10.1177/0265532215582282

de Ruiter, J. P., & Albert, S. (2017). An appeal for a methodological fusion of conversation analysis and experimental psychology. *Research on Language and Social Interaction, 50*(1), 90-107. doi:10.1080/08351813.2017.1262050

Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford: Oxford University Press.

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing, 11*, 125-144.

Drew, P. (1984). Speakers' reporting in invitation sequences. In J. Heritage & M. Atkinson (Eds.), *Structures of social action: Studies in conversation analysis* (pp. 129-151). Cambridge: Cambridge University Press.

Drew, P. (1997). 'Open' class repair initiators in response to sequential sources of trouble in conversation. *Journal of Pragmatics, 28*, 69-101.

Drew, P. (2018). Equivocal invitations (in English). *Journal of Pragmatics, 125*, 62-75.

Drew, P., & Heritage, J. (Eds.). (1992). *Talk at work: Interaction in institutional settings*. Cambridge, New York: Cambridge University Press.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt: Peter Lang.

Elder, C., Pill, J., Woodward-Kron, R., McNamara, T., Manias, E., Webb, G., & McColl, G. (2012). Health professionals' views of communication: Implications for assessing

performance on a health-specific English language test. *TESOL Quarterly, 46*(2), 409-419. doi:10.1002/tesq.26

Firth, A., & Wagner, J. (1997). On discourse, communication, and (some) fundamental concepts in SLA research. *The Modern Language Journal, 81*(3), 285-300.

Francis, D. (1989). Game identities and activities: Some ethnomethodological observations. In D. Crookall & D. Saunders (Eds.), *Communication and simulation* (pp. 53-68). Clevedon: Multilingual Matters.

Fulcher, G. (2003). *Testing second language speaking*. Harlow: Longman.

Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.

Fulcher, G. (2015). *Re-examining language testing: A philosophical and social inquiry.* CITY: Routledge.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5-29. doi:10.1177/0265532209359514

Gage, N. L. (1989). The paradigm wars and their aftermath: A "historical" sketch of research on teaching since 1989. *Educational Researcher, 18*(7), 4-10.

Galaczi, E. D. (2008). Peer-Peer Interaction in a Speaking Test: The Case of the First Certificate in English Examination. *Language Assessment Quarterly, 5*(2), 89-119.

Galaczi, E. D. (2014). Interactional Competence across Proficiency Levels: How do Learners Manage Interaction in Paired Speaking Tests? *Applied Linguistics (Oxford), 35*(5), 553-574.

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher- and lower-scoring students. *Language Testing, 27*(4), 585-602. doi:10.1177/0265532210364049

Gardner, R. C., & Wagner, J. (2004). *Second language conversations*. London: Continuum.

Garfinkel, H. (1967). *Studies in ethnomethodology*. Englewood Cliff, NJ: Prentice Hall.

Good, J. S., & Beach, W. A. (2005). Opening up gift-openings: Birthday parties as situated activity systems. *Text, 25*(5), 565-593.

Grabowski, K. (2013). Investigating the construct validity of a role-play test designed to measure grammatical and pragmatic knowledge at multiple proficiency levels. In S. J. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 149-171). New York: Palgrave Macmillan.

Greene, J. C. (2007). *Mixed methods in social inquiry*. San Francisco, CA: Jossey-Bass.

Greene, J. C. (2008). Is mixed methods social inquiry a distinctive methodology? *Journal of Mixed Methods Research, 2*(1), 7-22.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation design. *Educational Evaluation and Policy Analysis, 11*(3), 255-274.

Hall, J. K. (1995). "Aw, man, where you goin?": Classroom interaction and the development of L2 interactional competence. *Issues in Applied Linguistics, 6*(2), 37-62.

Hall, J. K., Hellermann, J., & Pekarek Doehler, S. (2011). *L2 interactional competence and development*. Bristol: Multilingual Matters.

He, A. W., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 1-24). Philadelphia: Benjamins.

Hellermann, J. (2007). The development of practices for action in classroom dyadic interaction: Focus on task openings. *The Modern Language Journal, 91*(1), 83-96.

Hellermann, J. (2008). *Social actions for classroom language learning*. Clevedon, UK: Multilingual Matters.

Hellermann, J. (2011). Members' methods, members' competencies: Looking for evidence of language learning in longitudinal investigation of other-initiated repair. In J. K. Hall, J. Hellerman, & S. Pekarek-Doehler (Eds.), *L2 interactional competence and development* (pp. 147-172). Bristol: Multilingual Matters.

Heritage, J. (1984a). A change-of-state token and aspects of its sequential placement. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action* (pp. 299–345). Cambridge: Cambridge University Press.

Heritage, J. (1984b). *Garfinkel and ethnomethodology*. Cambridge: Polity Press.

Heritage, J., & Sorjonen, M.-L. (1994). Constituting and maintaining activities across sequences: And-prefacing as a feature of question design. *Language in Society, 23*(1), 1-29.

Hudson, T. (2001). Indicators for pragmatic instruction: Some quantitative tools. In K. R. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 283-300). Cambridge: Cambridge University Press.

Hudson, T., Detmer, E., & Brown, J. D. (1992). *A framework for testing cross-cultural pragmatics*. (Technical Report # 2). University of Hawaii at Manoa, Second Language Teaching and Curriculum Center, Honolulu.

Hudson, T., Detmer, E., & Brown, J. D. (1995). *Developing prototypic measures of cross-cultural pragmatics*. (Technical Report # 7). University of Hawaii at Manoa, Second Language Teaching and Curriculum Center, Honolulu.

Huth, T. (2010). Can Talk Be Inconsequential? Social and Interactional Aspects of Elicited Second-Language Interaction. *The Modern Language Journal, 94*(4), 537-553. doi:10.1111/j.1540-4781.2010.01092.x

Huth, T. (2014). “When in Berlin…”: Teaching German telephone openings. *Die Unterrichtspraxis/Teaching German, 47*(2), 164-179.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). Harmondsworth: Penguin.

Jacoby, S., & Ochs, E. (1995). Co-construction: An introduction. *Research on Language and Social Interaction, 28*(3), 171-183.

Jang, E. E., Wagner, M., & Park, G. (2014). Mixed methods research in language testing and assessment. *Annual Review of Applied Linguistics, 34*, 123-153. doi:10.1017/s0267190514000063

Jefferson, G. (1978). Sequential aspects of storytelling in conversation. In J. Schenkein (Ed.), *Studies in the organization of conversational interaction* (pp. 219-248). New York, NY: Academic Press.

Jefferson, G. (1984a). On stepwise transition from talk about trouble to inaapropriately next-positioned matters. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social interaction: Studies in conversation analysis* (pp. 191-224). Cambridge: Cambridge University Press.

Jefferson, G. (1984b). On the organization of laughter in talk about troubles. In J. M. Atkinson & J. Heritage (Eds.), *Structure of social action* (pp. 346-369). Cambridge: Cambridge University Press.

Jefferson, G. (1988). On the sequential organization of troubles-talk in ordinary conversation. *Social Problems, 35*(4), 418-441.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research, 1*(2), 112-133. doi:10.1177/1558689806298224

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: Greewook Publishing.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kane, M. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice, 23*(2), 198-211.

Kane, M. (2017). Loosening psychometric constraints on educational assessment. *Assessment in Education: Principles, Policy & Practice, 24*(3), 447-453.

Kasper, G. (2006). When once is not enough: Politeness of multiple requests in oral proficiency interviews. *Multilingua - Journal of Cross-Cultural and Interlanguage Communication, 25*(3), 323-350. doi:10.1515/multi.2006.018

Kasper, G. (2009). Locating cognition in second language interaction and learning: Inside the skull or in public view? *IRAL - International Review of Applied Linguistics in Language Teaching, 47*(1). doi:10.1515/iral.2009.002

Kasper, G., & Ross, S. J. (2007). Multiple questions in oral proficiency interviews. *Journal of Pragmatics, 39*(11), 2045-2070. doi:10.1016/j.pragma.2007.07.011

Kasper, G., & Ross, S. J. (2013). Assessing second language pragmatics: An overview and introductions. In S. J. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 1-40). New York: Palgrave Macmillan.

Kasper, G., & Wagner, J. (2011). A conversation-analytic approach to second language acquisition. In D. Atkinson (Ed.), *Alternative approaches to second language acquisition* (pp. 117-142). New York: Routledge.

Kasper, G., & Youn, S. (2017). Transforming instruction to activity: Roleplay in language assessment. *Applied Linguistics Review*.

Kendrick, K. H., & Holler, J. (2017). Gaze direction signals response preference in conversation. *Research on Language and Social Interaction, 50*(1), 12-32. doi:10.1080/08351813.2017.1262120

Kim, H. (2018). What constitutes professional communication in aviation: Is language proficiency enough for testing purposes? *Language Testing, 35*(3), 403-426.

Kley, K. (2015). *Interactional competence in paired speaking tests: Role of paired task and test-taker speaking ability in co-constructed discourse.* (PhD Dissertation), University of Iowa, Iowa Research Online. Retrieved from http://ir.uiowa.edu/etd/1663

Knoch, U. (2010). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing, 16*, 81-96.

Knoch, U., & Chapelle, C. A. (2017). Validation of rating processes within an argument-based framework. *Language Testing*.

Knoch, U., McNamara, T. F., Woodward-Kron, R., Elder, C., Manias, E., Flynn, E., & Zhang, Y. (2015). Towards improved language assessment of written health professional communication: the case of the Occupational English Test. *Papers in Language Testing and Assessment, 4*(2), 60-66.

Kormos, J. (1999). Simulating conversations in oral-proficiency assessment: a conversation analysis of role plays and non-scripted interviews in language exams. *Language Testing, 16*(2), 163-188. doi:10.1177/026553229901600203

Koschmann, T. (2013). The perils of appropriation. *Qualitative Research in Psychology, 10*(240-243).

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal, 70*(4), 366-372.

Lantolf, J. P., & Frawley, W. (1985). Oral proficiency testing: A critical analysis. *Modern Language Journal, 69*(4), 337-345.

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests*. Cambridge: Cambridge University Press.

Lazaraton, A. (2008). Utilizing qualitative methods for assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., Vol. 7: Language testing and assessment, pp. 197-209). New York, NY: Springer.

Lazaraton, A. (2014). Spoken discourse. In A. Kunnan (Ed.), *Companion to language assessment* (pp. 1375-1389). New York: Wiley-Blackwell.

Lee, Y., & Greene, J. C. (2007). The predictive validity of an ESL placement test: A mixed methods approach. *Journal of Mixed Methods Research, 1*(4), 366-389.

Lee, Y.-A. (2006). Towards Respecification of Communicative Competence: Condition of L2 Instruction or its Objective? *Applied Linguistics, 27*(3), 349-376. doi:10.1093/applin/aml011

Lee, Y.-A., & Hellermann, J. (2014). Tracing developmental changes through conversation analysis: Cross-sectional and longitudinal analysis. *TESOL Quarterly, 48*(4), 763-788. doi:10.1002/tesq.149

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28*, 543-560.

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (1998a). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement, 2*(3), 266-283.

Linacre, J. M. (1998b). FACETS 3.17 Computer program. Chicago, IL: MESA Press.

Linacre, J. M. (2002). What do infit and outfit mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2006). FACETS Rasch measurement computer program (Version 3.61.0). Chicago: Winsteps.com.

Linacre, J. M. (2018). Facets computer program for many-facet Rasch measurement (Version version 3.81.0). Beaverton, Oregon: Winsteps.com.

Lindström, A., & Sorjonen, M.-L. (2013). Affiliation in conversation. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 350-369). Oxford, UK: Wiley-Blackwell.

Linn, R., Baker, E., & Dunbar, S. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Education Researcher, 20*(15-21).

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, Monograph Supplement, 3*, 635-694.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt: Peter Lang.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing, 12*, 54-71.

Mandelbaum, J. (2013). Storytelling in conversation. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 492-507). Oxford: Blackwell Publishing.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Maxwell, J. A., & Loomis, D. M. (2003). Mixed methods design: An alternative approach. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research*. Thousand Oaks, CA: Sage.

May, L. A. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly, 8*(2), 127-145. doi:10.1080/15434303.2011.565845

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

McNamara, T. F. (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics, 18*(4), 446-466.

McNamara, T. F. (2006). The challenge of Sam Messick's legacy. *Language Assessment Quarterly, 3*(1), 31-51.

McNamara, T. F., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell Publishing.

Mehan, H. (1979). *Learning lessons*. Cambridge, MA: Harvard University Press.

Mendoza, A., & Knoch, U. (2018). Examining the validity of an analytic rating scale for a Spanish test for academic purposes using the argument-based approach to validation. *Assessing Writing, 35*, 41-55.

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *American Educational Research Association, 18*(2), 5-11.

Messick, S. (1989b). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of interferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241-256.

Myford, C., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227.

Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing, 28*(4), 483-508. doi:10.1177/0265532211398110

Nakatsuhara, F. (2013). *The co-construction of conversation in group oral tests*. Frankfurt am Main: Peter Lang.

Nguyen, H. t. (2011). Achieving recipient design longitudinally: Evidence from a pharmacy intern in patient consultations. In J. K. Hall, J. Hellerman, & S. Pekarek-Doehler (Eds.), *L2 interactional competence development* (pp. 173-205). Bristol: Multilingual Matters.

Nguyen, H. t. (2012a). *Developing interactional competence: A conversation-analytic study of patient consultations in pharmacy*. London: Palgrave Macmillan.

Nguyen, H. t. (2012b). Social interaction and competence development: Learning the structural organization of a communicative practice. *Learning, Culture and Social Interaction, 1*, 127-142. doi:10.1016/j.lcsi.2012.05.006

Norris, J. M. (2001). Identifying rating criteria for task-based EAP assessment. In T. Hudson & J. D. Brown (Eds.), *A focus on language test development* (pp. 163-204). Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing, 19*, 277-295.

Ockey, G. J. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing, 26*(2), 161-186. doi:10.1177/0265532208101005

Ockey, G. J. (2014). The potential of the L2 group oral to elicit discourse with a mutual contingency pattern and afford equal speaking rights in an ESP context. *English for Specific Purposes, 35*, 17-29. doi:10.1016/j.esp.2013.11.003

Ockey, G. J., Koyama, D., Setoguchi, E., & Sun, A. (2015). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing, 32*(1), 39-62. doi:10.1177/0265532214538014

Okada, Y. (2010). Role-play in oral proficiency interviews: Interactive footing and interactional competencies. *Journal of Pragmatics, 42*, 1647-1668.

Okada, Y., & Greer, T. (2013). Pursuing a relevant response in oral proficiency interview role plays. In S. J. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 288-310). New York: Palgrave Macmillan.

Pekarek Doehler, S., & Berger, E. (2016). L2 interactional competence as increased ability for context-sensitive conduct: A longitudinal study of story-openings. *Applied Linguistics, 39*(4), 555-578.

Pekarek Doehler, S., & Pochon-Berger, E. (2011). Developing 'methods' for interaction: A cross-sectional study of disagreement sequences in French L2. In J. K. Hall, J. Hellerman, & S. Pekarek-Doehler (Eds.), *L2 interactional competence development* (pp. 206-243). Bristol: Multilingual Matters.

Pekarek Doehler, S., & Pochon-Berger, E. (2015). The development of L2 interactional competence: evidence from turn-taking organization, sequence organization, repair organization and preference organization. In T. Cadierno & S. W. Eskildsen (Eds.), *Usage-based perspectives on second language learning*. Berlin, Boston: De Gruyter Mouton.

Pekarek Doehler, S., & Wagner, J. (2010). *Analyzing change across time: Conceptual and methodological issues*. Paper presented at the International Conference on Conversation Analysis, Mannheim.

Pekarek Doehler, S., Wagner, J., & González-Martinez, E. (Eds.). (2018). *Longitudinal studies on the organization of social interaction*. London: Palegrave Macmillan.

Plough, I. (2018). Revisiting the speaking construct: The question of interactional competence. *Language Testing, 35*(3), 325-329.

Pomerantz, A. (1984). Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social actions: Studies in conversation analysis* (pp. 57-101). Cambridge: Cambridge University Press.

Rine, E. F., & Hall, J. K. (2011). Becoming the teacher: Changing Participant frameworks in international teaching assisstant discourse. In J. K. Hall, J. Hellerman, & S. Pekarek Doehler (Eds.), *L2 interactional competence development* (pp. 244-271). Bristol: Multilingual Matters.

Robinson, J. D. (2003). An interactional structure of medical activities during acute visits and its implications for patients' participation. *Health Communication, 15*(1), 27-57.

Robinson, J. D. (2013). Overall structural organization. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 257-280). West Sussex, UK: Wiley Blackwell.

Robinson, J. D., & Heritage, J. (2006). Physicians' opening questions and patients' satisfaction. *Patient Educ Couns, 60*(3), 279-285. doi:10.1016/j.pec.2005.11.009

Roever, C. (2011). Testing of second language pragmatics: Past and future. *Language Testing, 28*(4), 463-481. doi:10.1177/0265532210394633

Roever, C., & Kasper, G. (2018). Speaking in turns and sequences: Interactional competence as a target construct in testing speaking. *Language Testing, 35*(3), 331-355.

Ross, S. J. (2017). *Interviewing for language proficiency: Interaction and interpretation*. Basingstoke: Palgrave Macmillan.

Ross, S. J. (2018). Listener response as a facet of interactional competence. *Language Testing, 35*(3), 357-375.

Ross, S. J., & Kasper, G. (Eds.). (2013). *Assessing second language pragmatics*. Basingstoke: Palgrave Macmillan.

Sacks, H. (1972). On the analyzability of stories by children. In J. J. Gumperz & D. Hymes (Eds.), *Directions in sociolinguistics: The ethnography of communication* (pp. 325-345). New York: Rinehart & Winston.

Sacks, H. (1992). *Lectures on conversation* (Vol. 2). Oxford: Blackwell.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*, 696-735.

Sanders, R. E. (2003). Applying the skills concept to discourse and conversation: The remediation of performance defects in talk-in-interaction. In J. Greene & B. Burleson (Eds.), *The handbook of communication and social interaction skills* (pp. 221-256). Mahwah, NJ: Erlbaum.

Schegloff, E. A. (1968). Sequencing in conversational openings. *American Anthropologist, 70*(6), 1075-1095.

Schegloff, E. A. (1982a). Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In D. Tannen (Ed.), *Analyzing Discourse: Text and talk*. Washington D.C.: Georgetown University Press.

Schegloff, E. A. (1982b). Discourse as an interactional achievement: Some uses of ‘uh huh’ and other things that come between sentences. In D. Tannen (Ed.), *Analyzing discourse: Text and talk*. Georgetown: Georgetown University Press.

Schegloff, E. A. (1986). The routine as achievement. *Human Studies, 9*(2), 111-151.

Schegloff, E. A. (1988). Presequences and indirection: Applying speech act theory to ordinary conversation. *Journal of Pragmatics, 12*(1), 55-62.

Schegloff, E. A. (1990). On the organization of sequences as a source of "coherence" in talk-in-interaction. In B. Dorval (Ed.), *Conversational organization and its development* (pp. 51-77). Norwood, NJ: Ablex Publishing Co.

Schegloff, E. A. (1993). Reflections on quantification in the study of conversation. *Research on Language & Social Interaction, 26*(1), 99-128.

Schegloff, E. A. (1996). Confirming allusions: Toward an empirical account of action. *American Journal of Sociology, 102*(1), 161–216.

Schegloff, E. A. (2006). Interaction: The infrastructure for social institutions, the natural ecological niche for language, and the arena in which culture is enacted. In N. J. Enfield & S. C. Levinson (Eds.), *Roots of human society* (pp. 70-96). Oxford: Berg.

Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis*. Cambridge: Cambridge University Press.

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language, 53*(2), 361-382.

Schegloff, E. A., & Sacks, H. (1973). Opening up closings. *Semiotica, 8*(4), 289-327.

Seedhouse, P. (2012). What kind of interaction receives high and low ratings in Oral Proficiency Interviews? *English Profile Journal, 3*. doi:10.1017/s2041536212000025

Seedhouse, P. (2013). Oral prificiency interviews as varieties of interaction. In S. J. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 199-219). New York: Palgrave Macmillan.

Seedhouse, P., & Egbert, M. (2006). The interactional organisation of the IELTS Speaking Test *IELTS Research Reports 6* (pp. 161-206).

Seedhouse, P., & Nakatsuhara, F. (2018). *The discourse of the IELTS speaking test: Interactional design and practice*. Cambridge, UK: Cambridge University Press.

Sidnell, J., & Stivers, T. (Eds.). (2013). *The handbook of conversation analysis*. West Sussex, UK: Wiley-Blackwell.

Slater, S. J. (1980). Introduction to performance testing. In J. E. Spirer (Ed.), *Performance testing: Issues facing vocational education* (pp. 3-17). Columbus, OH: National Center for Research in Vocational Education.

Steensig, J. (2012). Conversation Analysis and Affiliation and Alignment *The Encyclopedia of Applied Linguistics*.

Stivers, T., Mondada, L., & Steensig, J. (2011). Knowledge, morality and affiliation in social interaction. In T. Stivers, L. Mondada, & J. Steensig (Eds.), *The morality of knowledge in conversation* (pp. 3-24). Cambridge: Cambridge University Press.

Tashakkori, A., & Teddlie, C. (2003). The past and the future of mixed methods research: From data triangulation to mixed methods in social and behavioral research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research*. Thousand Oaks, California: Sage.

Tebachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics* (Sixth Edition ed.). Boston: Pearson Educational.

Teddlie, C., & Tashakkori, A. (2010). *Sage handbook of mixed methods in social and behavioral research*. Thousand Oaks, CA: Sage Publications.

Teddlie, C., & Tashakkori, A. (2012). Common "core" characteristics of mized methods research: A review of critical issues and call for greater convergence. *American Behavioral Scientist, 56*(6), 774-788.

Theodórsdóttir, G. (2011). Second language interaction for business and learning. In J. K. Hall, J. Hellerman, & S. Pekarek Doehler (Eds.), *L2 interactional competence and development* (pp. 93-116). Bristol: Multilingual Matters.

Tominaga, W. (2013). The development of extended turns and storytelling in the Japanese oral prificiency interview. In S. J. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 220-257). New York: Palgrave Macmillan.

Van Lier, L. (1989). Reeling, writhing, drawing, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly, 23*, 489-508.

Veronis, E. M., & Gass, S. (1985). Non-native / non-native conversations: A model for negotiation of meaning. *Applied Linguistics, 6*, 71-90.

Walters, F. S. (2007). A conversation-analytic hermaneutic rating protocol to assess L2 oral pragmatic competence. *Language Testing, 24*(2), 155-183.

Waring, H. Z. (2018). Teaching L2 interactional competence: Problems and possibilities. *Classroom Discourse, 9*(1), 57-67.

Watson, D. R., & Sharrock, W. W. (1990). Realities in simulation/gaming. In D. Crookall & R. Oxford (Eds.), *Simulation, gaming and language learning* (pp. 231-238). New York: Newbury House.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*, 199-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Wong, J., & Waring, H. Z. (2010). *Conversation analysis and second language pedagogy: A guide for ESL/EFL Teachers*. New York: Routledge.

Yamashita, S. (2008). Investgating interlanguage pragmatic ability: What are we testing? In E. Alcon & A. Martinez-Flor (Eds.), *Investigating pragmatics in foreign language learning, teaching and testing* (pp. 201-223). Bristol: Multilingual Matters.

Youn, S. (2013). *Validating task-based assessment of l2 pragmatics in interaction using mixed methods*. Doctoral dissertation. Department of Second Language Studies. University of Hawai'i at Manoa.

Youn, S. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing, 32*(2), 199-225. doi:10.1177/0265532214557113

Young, R. (2000). *Interactional competence: Challenges for validity*. Paper presented at the Annual meeting of the AAAL and the LTRC, Vancouver, Canada.

Young, R. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 426-433). London & New York: Routledge.

Young, R., & He, A. W. (1998). *Talking and testing: Discourse approaches to the assessment of oral proficiency*. Amsterdam: John Benjamins.

Young, R., & Milanovic, M. (1992). Discourse variation in oral proficiency interviews. *Studies in Second Language Aquisition, 14*, 403-424.

APPENDIX A

Excerpt from the Socializing Unit textbook

8. Mr Syms and Cathy are talking about their plans. Listen and complete Cathy's diary.

12 AUGUST	12 AUGUST
9:00	16:00
10:00	17:00
11:00 11:40 Mr Syms arrives at Warsaw airport, flight BA120	18:00
12:00 12 (approx) - 11 Casolare	19:00
13:00	20:00
14:00	21:00
15:00	22:00

9. Listen again and complete the sentences from the dialogue.

1. It's _____now. We'll be _____in five minutes.
2. I thought you might like to _____your hotel first and_____ _____your things.
3. Then we _____ a spot of lunch. There's nice Italian place _____ your hotel.
4. After that we _____to the office.
5. We _____with the sales team at two, as you know.
6. At four we _____the production plant.
7. That _____an hour.
8. Then perhaps you _____ a taxi back to your hotel and _____ for a bit.
9. I _____ again at about seven for dinner.
10. It _____really good. We _____ to this fantastic French restaurant.

11

Language Box: Talking about Plans

Grammar focus

There are many ways to talk about future plans in English, and often you can say the same thing in different ways. Here are some ways to talk about plans:

- **using modals verbs such as can, could, might, should, etc:**

  I thought you might like to check into your hotel first.
  Then we can go to the office.
  That should only take an hour.

- **using the present tense:**

  We have the meeting with the sales team at two.
  It's the big company dinner tonight.

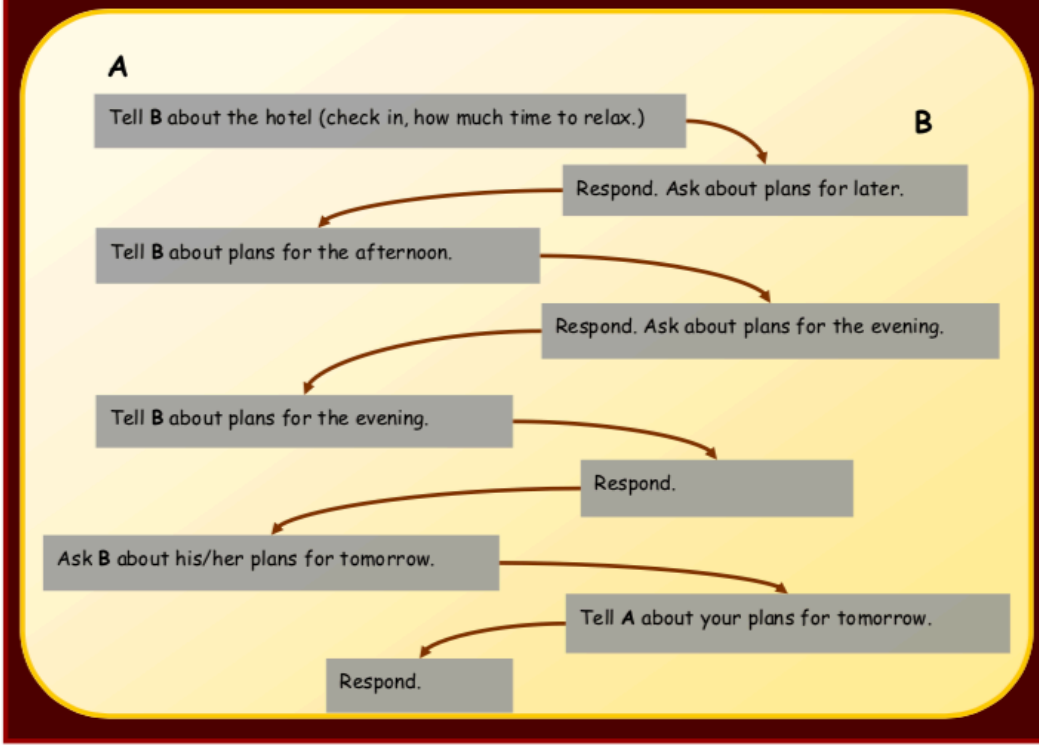
- **using will:**

  I'll pick you up again at about seven for dinner.

- **using going to:**

  At four we're going to visit the production plant.
  After that we're going to this fantastic French restaurant.

10. Work with a partner to make a dialogue. Person A: you are the host. You are dropping B off at his/her hotel. Person B: you are the visitor.



12

APPENDIX B

FACETS Student Measurement Report

Students	Measure	Model *S.E.*	Infit MnSq	Outfit MnSq
1	0.78	0.14	0.71	0.75
2	-0.04	0.12	0.97	0.94
3	0.82	0.14	0.82	0.79
4	0.71	0.13	1.35	1.21
5	0.04	0.12	0.66	0.64
6	0.55	0.13	0.91	0.82
7	0.64	0.13	0.68	0.66
8	0.29	0.12	1.21	1.27
9	0.82	0.14	1.02	0.9
10	0.15	0.12	0.7	0.7
11	1.02	0.15	1.05	0.97
12	1.33	0.16	0.77	0.78
13	1.16	0.15	1.13	1.38
14	1.48	0.18	0.94	1.05
15	0.35	0.12	1.5	1.56
16	0.75	0.14	0.8	0.91
17	0.38	0.12	1.25	1.25
18	0.11	0.12	0.73	0.79
19	0.27	0.12	0.92	0.98
20	0.9	0.14	1.14	1.23
21	0.64	0.13	0.88	0.88
22	0.98	0.14	0.76	0.85
23	1.18	0.16	0.84	0.78
24	0.71	0.13	1	0.93
25	0.86	0.14	0.84	0.87
26	0.5	0.13	0.91	0.94
27	0.75	0.14	1.48	1.4
28	0.3	0.12	1.14	1.18
29	0.2	0.12	0.8	0.88
30	1	0.15	1.36	1.5
31	0.44	0.13	1.33	1.3
32	0.29	0.12	1.54	1.58
33	0.44	0.13	1.79	1.76
34	0.82	0.14	1.45	1.43
35	-0.11	0.12	0.51	0.51
36	0.1	0.12	0.62	0.62
37	-0.06	0.12	0.98	0.99
38	0.11	0.12	1.04	1.02
39	0.71	0.13	0.83	0.83
40	1.07	0.15	0.88	1.06
42	0.59	0.13	1.61	1.54
43	0.49	0.13	0.98	1.05
44	0.54	0.13	1.07	1.16
45	0.62	0.13	1.32	1.38
46	0.42	0.13	1.23	1.05
47	-0.12	0.12	1.05	1.08
48	0.96	0.14	1.8	1.84
49	1.07	0.15	1.42	0.98

Students	Measure	Model *S.E.*	Infit MnSq	Outfit MnSq
50	0.6	0.13	0.99	0.86
51	-0.11	0.12	0.83	0.78
52	-0.26	0.11	0.62	0.57
53	-0.43	0.11	0.61	0.57
54	0.67	0.13	1.11	0.94
55	0.41	0.13	0.69	0.7
56	0.86	0.14	0.91	0.76
57	0	0.12	0.86	0.82
58	0.2	0.12	0.69	0.68
59	1.23	0.16	1.76	1.5
60	0.66	0.13	1.17	1.03
61	-0.51	0.11	0.87	1.14
62	0.01	0.12	0.97	0.98
63	0.5	0.13	1.09	1.03
64	0.27	0.12	1.56	1.68
65	0.82	0.14	0.91	0.81
66	1.16	0.15	1.9	2.51
67	1.45	0.17	1.36	2
68	0.33	0.12	1.65	1.74
69	1	0.15	1.61	1.21
70	0.54	0.13	1.47	1.46
71	0.23	0.12	0.98	0.99
72	0.26	0.12	1.11	1.11
73	0.11	0.12	1.05	1
74	0.15	0.12	0.96	0.94
75	0.04	0.12	1.03	0.95
76	0.38	0.12	0.67	0.7
77	0.17	0.12	0.77	0.78
78	0.32	0.12	0.78	0.75
79	0.5	0.13	1.04	0.96
80	0.44	0.13	0.92	0.88
81	0.54	0.13	1.1	1.08
82	0.46	0.13	0.9	0.86
83	0.23	0.12	1.68	1.61
84	0.6	0.13	1.01	0.97
85	0.8	0.14	0.97	0.97
86	0.49	0.13	1.06	0.96
87	0.41	0.13	0.68	0.64
88	0.76	0.14	1.06	0.86
89	-0.02	0.12	0.86	0.87
90	0.05	0.12	1.09	1.13
91	0.59	0.13	0.96	0.97
92	1.25	0.16	0.55	0.63
93	0.96	0.14	0.48	0.52
94	0.64	0.13	0.92	0.95
97	0.94	0.14	0.88	0.83
103	-0.19	0.12	1.16	1.12
104	-0.06	0.12	1.38	1.37
105	0.24	0.12	1.22	1.31
106	0.27	0.12	1.11	1.18
107	0.3	0.12	1.17	1.29
108	0.71	0.13	1.11	1.32
109	0.6	0.13	1.35	1.51
110	1	0.15	1.35	1.46

Students	Measure	Model *S.E.*	Infit MnSq	Outfit MnSq
111	0.73	0.13	1.07	1.13
112	0.39	0.13	1.71	1.73
113	0.39	0.13	1.39	1.39
114	0.39	0.13	1.47	1.57
115	1.25	0.16	1.55	1.09
116	1	0.15	1.22	0.99
117	1.61	0.19	0.97	0.89
118	0.23	0.12	0.76	0.74
119	0.05	0.12	1.07	1.22
120	0.26	0.12	0.7	0.8
121	0.07	0.12	1.03	1.2
122	0.15	0.12	0.82	0.89
123	0.11	0.12	0.88	0.95
125	0.03	0.12	0.62	0.59
126	0.2	0.12	0.89	0.86
127	-0.26	0.11	0.71	0.69
130	-0.23	0.11	0.71	0.72
131	1.04	0.15	0.59	0.64
132	0.94	0.14	0.53	0.59
133	0.84	0.14	1.01	1.31
134	1.2	0.16	0.84	1.01
135	1.28	0.16	0.62	0.67
136	-0.16	0.12	0.78	0.8
137	0.11	0.12	0.67	0.7
138	-0.15	0.12	0.72	0.74
139	-0.23	0.11	0.67	0.74
140	-0.06	0.12	0.69	0.69
142	0.52	0.13	1.1	1.17
143	0.44	0.13	1.39	1.44
145	0.96	0.14	1.3	1.31
146	0.59	0.13	1.04	1.13
147	0.92	0.14	1.35	1.32
149	0.52	0.13	0.59	0.7
150	0.54	0.13	0.74	0.86
152	0.71	0.13	0.68	0.73
153	1.09	0.15	1.07	0.96
160	0.32	0.12	0.91	0.92
161	0.66	0.13	1.58	1.75
162	0.23	0.12	1.22	1.19
163	0.15	0.12	1.01	0.97
165	0.41	0.13	0.72	0.76
167	1.18	0.16	1.79	1.24
177	0.24	0.12	1.11	1.25
178	0.18	0.12	0.94	0.99
179	0.41	0.13	1.21	1.39
180	0.11	0.12	0.75	0.83
M	0.49	0.13	1.04	1.05
SD	0.43	0.02	0.31	0.33

Note. Reliability = 0.91; Separation index: 3.11; Fixed chi-square: 1474.9 (*d.f.*=147; p<.00), RMSEA = 0.13